YAYASAN PRIMA AGUS TEKNIK

Aplikasi Geografis Dengan Komputasi Big Data

oleh:
Dr. Joseph Teguh Santoso, S.Kom, M.Kom

Aplikasi Geografis dengan Komputasi Big Data

Penulis :
Dr. Joseph Teguh Santoso, S.Kom., M.Kom

ISBN : 9 786238 120802

Editor :
Muhammad Sholikan, M.Kom

Penyunting :
Dr. Mars Caroline Wibowo. S.T., M.Mm.Tech

Desain Sampul dan Tata Letak :
Irdha Yunianto, S.Ds., M.Kom

Penebit :
Yayasan Prima Agus Teknik Bekerja sama dengan
Universitas Sains & Teknologi Komputer (Universitas STEKOM)

Anggota IKAPI No: 279 / ALB / JTE / 2023

Redaksi :
Jl. Majapahit no 605 Semarang
Telp. (024) 6723456
Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

Distributor Tunggal :
Universitas STEKOM
Jl. Majapahit no 605 Semarang
Telp. (024) 6723456
Fax. 024-6710144
Email : info@stekom.ac.id

KATA PENGANTAR

Puji syukur pada Tuhan Yang Maha Esa bahwa buku yang berjudul " Aplikasi Geografis dengan Komputasi Big Data " telah dapat diselesaikan dengan baik. Dalam era digital yang terus berkembang, integrasi antara Aplikasi Geografis (Geographic Information System/GIS) dan Komputasi Big Data telah membuka pintu menuju pemahaman yang lebih mendalam terhadap lingkungan dan fenomena geografis. Kombinasi kekuatan GIS sebagai alat analisis spasial dan kemampuan Komputasi Big Data untuk mengelola, memproses, dan menganalisis volume data yang besar membawa dampak positif yang luar biasa pada pemodelan, prediksi, dan pengambilan keputusan di berbagai bidang.

Aplikasi Geografis dengan pendekatan Big Data memungkinkan kita untuk mengeksplorasi dan memahami kompleksitas pola spasial serta dinamika geografis dengan tingkat detail yang sebelumnya sulit dicapai. Dengan adopsi teknologi ini, kita dapat merinci informasi geografis menjadi wawasan yang lebih mendalam, memfasilitasi perencanaan kota yang berkelanjutan, pemantauan lingkungan, serta manajemen risiko bencana secara lebih efektif.

Melalui buku ini, kita akan menjelajahi peran integral Aplikasi Geografis dalam konteks Komputasi Big Data, menggali potensi sinergi keduanya untuk memecahkan tantangan kompleks yang dihadapi oleh masyarakat modern. Dengan menggabungkan kekuatan GIS dan kemampuan analisis Big Data, kita dapat menciptakan solusi inovatif yang membantu memahami, merencanakan, dan mengelola lingkungan geografis dengan lebih canggih dan berdaya saing.

Buku ini dibagi menjadi 11 bab. Bab 1 buku ini akan menjelaskan pengantar big data untuk aplikasi geografis dan menerangkan tentang pemecahan masalah dan pengetahuan geografis, menangani data geografis yang sangat besar dan menganalisis serta memvisualisasikan data geografis. Bab 2 buku ini akan membahas tentang pendekatan D_ELT untuk mentransformasi dan menganalisis data secara efisien terutama data besar geografis. Bab 3 buku ini akan menerangkan tentang komputasi geografis dan merancang algoritma analisis overlay paralel berkinerja tinggi dengan mempertimbangkan kompleksitas bentuk polygon. Bab 4 buku ini mempelajari dampak distribusi grafis spasial, penyimpanan data spasial, dan metode pengindeksan terhadap efisiensi analisis overlay. Bab 5 buku ini membahas tentang platform GEE berkinerja tinggi dengan akurasi yang setara dengan GIS tradisional.

Bab 6 buku ini akan membahas tentang pemetaan berbasis web interaktif dikembangkan untuk mengintegrasikan metode pembelajaran mesin tingkat lanjut dengan visualisasi spesifik untuk mengkarakterisasi pola perilaku perjalanan angkutan umum dan untuk memungkinkan eksplorasi visual pola mobilitas angkutan umum pada skala dan resolusi berbeda dalam ruang dan waktu. Bab 7 buku ini memperkenalkan metode pembelajaran mendalam untuk mengekstrak informasi emosional masyarakat yang terperinci dari data

besar media sosial Tiongkok untuk membantu dalam analisis bencana. Bab 8 buku ini membahas tentang mengembangkan metode baru untuk mengekstraksi jalan raya yang hilang dengan merekonstruksi topologi jalan dari data lintasan navigasi seluler yang besar. bab selanjutnya, Bab 9 akan menjelaskan tentang mengabstraksi permasalahan geografis sebagai suatu tugas yang selanjutnya dapat diuraikan menjadi beberapa subtugas.

Bab 10 buku ini merancang model konseptual yang disebut GeoKG berdasarkan enam elemen di sekitar pertanyaan geografis, kemudian melengkapi operator konstruksi DL dan akhirnya memberikan formalisasi model dengan operator tersebut. Bab 11 sekaligus menjadi bab penutup buku ini. Pada bab ini membahas metode dan teknik untuk memecahkan masalah yang saling terkait yang dihadapi saat mentransmisikan, memproses, dan menyajikan metadata untuk data Observasi dan Pemodelan Sistem Bumi (ESOM) yang heterogen. Akhir kata semoga buku ini berguna bagi para pembaca. Terima kasih.

Semarang, Januari 2024
Penulis

Dr. Joseph Teguh Santoso, S.Kom, M.Kom

DAFTAR ISI

BAB 1
PENGANTAR KOMPUTASI BIG DATA UNTUK APLIKASI GEOGRAFIS

Konvergensi big data dan komputasi geografis telah membawa tantangan dan peluang bagi GIScience sehubungan dengan pengelolaan, pemrosesan, analisis, pemodelan, dan visualisasi data geografis. Edisi khusus ini menyoroti kemajuan terkini dalam mengintegrasikan pendekatan komputasi baru, metode spasial, dan strategi pengelolaan data untuk mengatasi tantangan big data Geografis dan sekaligus menunjukkan peluang penggunaan big data untuk aplikasi geografis. Kemajuan penting yang disoroti di sini adalah integrasi pemikiran komputasi dan pemikiran spasial serta transformasi ide dan model abstrak menjadi struktur data dan algoritma yang konkret. Buku ini pertama-tama memperkenalkan latar belakang dan motivasi terbitan khusus ini, diikuti dengan tinjauan sepuluh artikel yang disertakan. Kesimpulan dan arah penelitian masa depan disediakan di bagian terakhir.

1.1 PENDAHULUAN

Sistem observasi bumi dan simulasi model menghasilkan data geografis yang berbeda, dinamis, dan tersebar secara geografis dalam jumlah besar dengan resolusi spatiotemporal yang semakin halus. Sementara itu, perangkat pintar, sensor berbasis lokasi, dan platform media sosial yang tersebar luas memberikan informasi geografis yang luas tentang aktivitas kehidupan sehari-hari. Menganalisis aliran data besar geografis secara efisien memungkinkan kita menyelidiki pola kompleks dan mengembangkan sistem pendukung keputusan baru, sehingga memberikan nilai yang belum pernah ada sebelumnya bagi sains, teknik, dan bisnis. Namun, menangani lima "V" (*Volume* – volume, *Variety* – variasi, *Velocity* – kecepatan, *Validity* – kebenaran, dan *Value* – nilai) big data geografis merupakan tugas yang menantang karena sering kali perlu diproses, dianalisis, dan divisualisasikan dalam konteks ruang dan waktu yang dinamis.

Menyusul serangkaian sesi sukses yang diselenggarakan pada Pertemuan Tahunan American Association of Geographers (AAG) sejak tahun 2015, edisi khusus tentang "Komputasi Big Data untuk Aplikasi Geografis" oleh ISPRS International Journal of Geo-Information bertujuan untuk menangkap upaya terbaru dalam memanfaatkan, mengadaptasi, dan mengembangkan pendekatan komputasi baru, metode spasial, dan strategi pengelolaan data untuk mengatasi tantangan big data geografis untuk mendukung aplikasi di berbagai domain, seperti perubahan iklim, manajemen bencana, dinamika manusia, kesehatan masyarakat, serta lingkungan dan teknik.

Secara khusus, buku ini bertujuan untuk membahas topik-topik penting berikut: (1) infrastruktur geo-siber yang mengintegrasikan prinsip-prinsip spatiotemporal dan teknologi komputasi tingkat lanjut (misalnya, GPU (komputasi unit pemrosesan grafis), komputasi multicore, komputasi kinerja tinggi, dan komputasi awan); (2) inovasi dalam pengembangan kerangka dan arsitektur komputasi dan pemrograman (misalnya MapReduce, Spark) atau algoritma komputasi paralel untuk aplikasi geografis; (3) strategi pengelolaan dan model

penyimpanan data geografis baru yang dipadukan dengan komputasi berkinerja tinggi untuk kueri, pengambilan, dan pemrosesan data yang efisien (misalnya, mekanisme pengindeksan spatiotemporal baru); (4) metode komputasi baru yang mempertimbangkan kolokasi spatiotemporal (lokasi dan hubungan) pengguna, data, dan sumber daya komputasi; (5) metode pemrosesan, penambangan, dan visualisasi data besar geografis menggunakan komputasi kinerja tinggi dan kecerdasan buatan; (6) mengintegrasikan alur kerja ilmiah dalam komputasi awan dan/atau lingkungan komputasi berkinerja tinggi; dan (7) penelitian, pengembangan, pendidikan, dan visi lainnya terkait komputasi data besar geografis. Editorial ini memberikan ringkasan dari sepuluh artikel yang termasuk dalam terbitan ini dan menyarankan arah penelitian masa depan dalam bidang ini berdasarkan pengamatan kolektif kami.

Pembahasan dalam buku ini memberikan kontribusi yang signifikan terhadap penggunaan komputasi data besar untuk mengatasi berbagai masalah geografis (mulai dari mobilitas manusia hingga manajemen bencana hingga penemuan pengetahuan) dengan menggabungkan metodologi, struktur data, dan algoritme baru dengan kerangka komputasi tingkat lanjut (dari geoanalitik visual, pembelajaran mendalam pada komputasi awan, dan MapReduce/Spark). Dengan menggunakan sepuluh sumber data besar yang berbeda (misalnya media sosial, penginderaan jarak jauh, dan Internet of Things), isu ini menunjukkan nilai dan pentingnya mengintegrasikan pendekatan komputasi dan metode geografis dalam memajukan penemuan ilmiah dan aplikasi domain (Tabel 1.1).

Tabel 1.1. Ringkasan pendekatan big data dan komputasi geografis

Kategori	Aplikasi Geografis	Sumber Data Besar	Pendekatan Komputasi
Metode Komputasi Big Data	Pemrosesan awal data geografis	Data sensor melalui Internet of Things (IoT)	Ekstraksi paralel, transformasi, pemuatan, MapReduce/Hadoop
	Analisis hamparan	Penggunaan lahan (sebagai studi kasus)	Komputasi kinerja tinggi dengan Spark, komputasi awan
	Prediksi perubahan penggunaan lahan	Penginderaan jauh (Landsat)	Pemodelan paralel dengan MapReduce/Hadoop, komputasi awan
	Analisis medan skala global	Ketinggian global	Mesin Google Earth, komputasi awan
Penambangan Data Besar	Mobilitas manusia (penemuan pola)	Angkutan umum	Pembelajaran mesin (algoritma pengelompokan), analisis visual
	Penanggulangan bencana (mitigasi gempa bumi)	Media sosial	Pembelajaran mendalam (CNN), analisis spasialtemporal
	Generasi jalan hilang	Navigasi (lintasan)	Satu set algoritma komputasi baru

Representasi Pengetahuan	Pemecahan masalah geografis	Data heterogen melalui layanan online	Alur kerja, geoproses online, basis pengetahuan
	Representasi pengetahuan geografis	Ontologis	Grafik pengetahuan, ontologi
Pencarian Data Besar	Pengelolaan dan pencarian big data geografis (data iklim)	Iklim	Katalog berbasis infrastruktur siber, pengindeksan spatiotemporal

1.2 METODE KOMPUTASI BIG DATA

Pemrosesan dan analisis data geografis, seperti transformasi geometri, konversi sistem referensi koordinasi, dan evaluasi hubungan spasial, sering kali mencakup sejumlah besar perhitungan aritmatika floating-point. Sejalan dengan itu, kerangka kerja dan sistem berbasis MapReduce dan Spark, seperti SpatialHadoop dan GeoSpark, dikembangkan untuk mempercepat komputasi ini. Selain itu, platform komputasi berbasis awan, seperti Google Earth Engine (GEE) untuk data observasi bumi besar, semakin banyak digunakan dalam studi dan aplikasi geografis. Untuk mengoptimalkan kinerja algoritma paralel untuk pemrosesan, analisis, atau pemodelan geografis ketika menggunakan kerangka tujuan umum tersebut, karakteristik spasial dari data dan algoritma harus dipertimbangkan untuk desain algoritma. Dalam buku ini berfokus pada komputasi paralel dan menyoroti adaptasi kerangka komputasi yang ada untuk prapemrosesan data geografis, desain algoritma paralel, pemodelan simulasi, dan analisis data.

Seringkali diperlukan waktu lama untuk mempersiapkan kumpulan data geografis untuk sistem komputasi data ini, yang umumnya melibatkan proses ekstraksi, transformasi, dan pemuatan (yaitu, ETL). Untuk menangani data besar dalam proses ETL, Jo dan Lee mengusulkan metode baru, D_ELT (delayedextracting–loading–transforming), untuk mengurangi waktu yang diperlukan untuk transformasi data dalam platform Hadoop dengan memanfaatkan paralelisasi berbasis MapReduce. Dengan menggunakan data sensor besar dengan berbagai ukuran dan analisis geografis dengan tingkat kompleksitas yang bervariasi, beberapa eksperimen dilakukan untuk mengukur kinerja keseluruhan sistem D_ELT, ETL tradisional, dan ekstraksi–pemuatan–transformasi (ELT). Hasilnya menunjukkan bahwa D_ELT mengungguli ETL dan ELT. Selain itu, semakin besar jumlah data atau semakin tinggi kompleksitas analisis, semakin baik kinerja D_ELT dibandingkan pendekatan ETL dan ELT tradisional.

Zhao dkk. merancang algoritma paralel untuk analisis overlay, yang menggunakan pengukuran kompleksitas bentuk poligon sebagai faktor kunci untuk partisi data dalam kombinasi dengan indeks spasial terdistribusi dan filter persegi batas minimum. Algoritma paralel diimplementasikan berdasarkan Spark, kerangka komputasi terdistribusi yang banyak digunakan untuk aplikasi skala besar. Hasil percobaan menunjukkan partisi data berdasarkan kompleksitas bentuk secara efektif meningkatkan penyeimbangan beban di antara beberapa node komputasi, sehingga menghasilkan efisiensi komputasi dari algoritma paralel. Buku ini

menunjukkan bahwa definisi dan pengukuran yang tepat dari properti data dan/atau algoritma (tidak peduli betapa sederhananya mereka) untuk mencerminkan intensitas komputasi adalah hal yang sangat penting untuk peningkatan kinerja algoritma paralel.

Model CA-Markov adalah salah satu model extended cell automata (CA) yang paling banyak digunakan dan telah digunakan dalam prediksi dan simulasi perubahan penggunaan lahan. Karena simulasi dan prediksi perubahan penggunaan lahan melibatkan sejumlah besar data dan perhitungan, banyak algoritma CA paralel telah dirancang untuk mensimulasikan pertumbuhan perkotaan berdasarkan berbagai model komputasi, termasuk unit pemrosesan pusat (CPU) dan GPU. Meskipun metode CA paralel menggabungkan hubungan spasial antar sel, metode ini tidak dapat mempertahankan hubungan antar partisi setelah area studi dibagi menjadi beberapa bagian, sehingga menghasilkan hasil prediksi yang berbeda. Sementara itu, metode Markov tradisional dapat menjaga integritas seluruh wilayah studi namun kurang mampu menggabungkan hubungan spasial antar sel. Alternatifnya, kerangka kerja MapReduce mampu melakukan pemrosesan paralel secara efisien bila digabungkan dengan model CA-Markov; masalah utama dalam segmentasi dan pemeliharaan koneksi spasial masih belum terselesaikan. Dengan demikian solusi berbasis MapReduce untuk meningkatkan model paralel CA-Markov untuk prediksi perubahan penggunaan lahan. Hasilnya menunjukkan bahwa model CA-Markov paralel tidak hanya memecahkan paradoks bahwa model CA-Markov tradisional tidak dapat secara bersamaan mencapai integritas dan segmentasi untuk simulasi dan prediksi perubahan penggunaan lahan, namun juga mencapai efisiensi dan akurasi.

Pengembangan algoritma analisis medan berdasarkan GEE (disebut TAGEE) untuk menghitung berbagai atribut medan, misalnya kemiringan, aspek, dan kelengkungan, untuk resolusi dan luas geografis yang berbeda.

Dengan menggunakan geometri bola yang diukur dengan jarak lingkaran besar, TAGEE tidak memerlukan data masukan DEM untuk diproyeksikan pada bidang datar. Eksperimen menunjukkan bahwa TAGEE dapat memberikan hasil yang serupa bila dibandingkan dengan paket perangkat lunak GIS konvensional. Dengan memanfaatkan kapasitas komputasi GEE berkinerja tinggi, TAGEE mampu secara efisien menghasilkan serangkaian produk atribut medan pada resolusi spasial apa pun dalam skala global. Buku ini mewakili paradigma komputasi geografis yang muncul di era data besar. Seiring dengan semakin matangnya platform komputasi awan seperti GEE, komputasi geografis tidak lagi dibatasi oleh sumber daya komputasi dan kumpulan data yang tersedia secara lokal. Penerapan algoritma/model geografis yang kompleks pada resolusi spasial tinggi dan skala global telah terlihat dalam beberapa tahun terakhir dan akan segera menjadi hal yang biasa.

1.3 PENAMBANGAN DATA BESAR

Penginderaan sosial, di mana manusia mewakili jaringan sensor yang besar, telah muncul sebagai pendekatan pengumpulan data baru di era big data. Tiga penelitian berikut oleh Zhang et al, Yang dkk, dan Wu dkk. menunjukkan kekuatan mengintegrasikan data penginderaan sosial (transportasi umum, media sosial, dan telepon seluler) dan teknik

komputasi data besar untuk mendukung aplikasi geografis termasuk mobilitas manusia, manajemen bencana, dan transportasi.

Zhang dkk. mengembangkan pendekatan baru untuk menambang dan memvisualisasikan pola mobilitas manusia dari data angkutan umum besar multisumber, yang bertujuan untuk mendukung perencanaan dan pengelolaan transportasi dengan memberikan pemahaman yang lebih baik tentang pola pergerakan manusia dalam ruang dan waktu. Untuk mengekstrak pola perjalanan secara efisien dari sumber data yang sangat heterogen, penelitian ini mengembangkan algoritme pengelompokan untuk mengekstraksi koridor transit yang menunjukkan hubungan antara wilayah berbeda dan algoritme penyematan grafik untuk mengungkap struktur komunitas mobilitas hierarkis. Selain algoritma pembelajaran mesin baru, karya ini juga menyediakan sistem analisis geo-visual berbasis web yang dapat diskalakan termasuk teknik visualisasi untuk memungkinkan pengguna menjelajahi pola yang diekstraksi secara interaktif. Sistem ini dievaluasi oleh 23 pengguna dengan latar belakang berbeda dan hasilnya mengkonfirmasi kegunaan dan efisiensi pendekatan analisis geo-visual terintegrasi untuk penemuan pola pergerakan manusia dari data besar angkutan umum. Buku ini menunjukkan kekuatan mengintegrasikan big data geografis, algoritma pembelajaran mesin, dan pendekatan analisis geo-visual untuk mendukung aplikasi transportasi.

Yang dkk. memperkenalkan metode pembelajaran mendalam untuk secara efisien melakukan analisis sentimen terhadap data media sosial yang besar untuk membantu mitigasi bencana. Pekerjaan ini merancang kerangka lima fase untuk ekstraksi otomatis emosi publik dari data mikro-blog Sina yang diberi geotag termasuk pengumpulan dan pemrosesan data, klasifikasi emosi, dan analisis spatiotemporal. Untuk mengklasifikasikan emosi (takut, cemas, sedih, marah, netral, dan positif), model jaringan saraf konvolusional (CNN) dirancang dan dilatih dengan mengubah teks mentah menjadi vektor kata. Untuk menunjukkan efisiensi pendekatan ini, gempa bumi di Ya'an, Tiongkok, pada tahun 2013 digunakan sebagai studi kasus. Berdasarkan model yang dilatih, emosi masyarakat di wilayah studi diklasifikasikan pada periode waktu yang berbeda setelah gempa bumi. Analisis spatiotemporal kemudian dilakukan untuk mengkaji dinamika sentimen masyarakat terhadap gempa bumi dalam ruang dan waktu. Hasil penelitian menunjukkan bahwa pendekatan yang diusulkan secara akurat mengklasifikasikan emosi dari data media sosial yang besar (>81%), memberikan informasi emosional publik yang berharga untuk mitigasi bencana.

Wu dkk. mengusulkan pendekatan tiga langkah untuk mendeteksi segmen jalan yang hilang dari data navigasi berbasis ponsel di lingkungan perkotaan. Langkah pertama mereka adalah menerapkan pemfilteran pada data navigasi untuk menghapus data yang terkait dengan pergerakan pejalan kaki dan segmen jalan yang ada. Kemudian, sebagai langkah kedua, garis tengah jalan yang hilang dibuat menggunakan algoritma clustering. Membangun topologi jalan yang hilang dan menghubungkan jalan yang terdeteksi dengan jaringan jalan yang ada adalah langkah ketiga. Wu dkk. menerapkan pendekatan ini di wilayah studi (sekitar 6 kilometer persegi) di Shanghai, Cina. Berdasarkan ~10 juta titik GPS yang dikumpulkan dari navigasi seluler pada tahun 2017, penelitian ini mengevaluasi kemampuan pendekatan tiga

langkah mereka dalam mendeteksi jalan yang hilang. Hasilnya menunjukkan kinerja pendekatan tiga langkah ini berdasarkan data ponsel, menyadari tantangan komputasi dari pendekatan mereka ketika berhadapan dengan kumpulan data yang lebih besar.

1.4 REPRESENTASI PENGETAHUAN

Zhuang dkk. membahas masalah yang belum diteliti, yaitu representasi dan berbagi pengetahuan terkait penyelesaian masalah geografis. Melalui proses abstraksi dan dekomposisi, karya ini mendekonstruksi masalah geografis menjadi tugas-tugas yang beroperasi pada tiga rincian berbeda. Selain deskripsi tingkat tinggi, pekerjaan ini memformalkan proses penyelesaian masalah geografis menjadi basis pengetahuan dengan menciptakan serangkaian ontologi untuk tugas, proses, dan operasi GIS. Dengan menggunakan analisis peringatan dini meteorologi sebagai studi kasus, penelitian ini berhasil menunjukkan bagaimana menangkap pengetahuan abstrak pemecahan masalah geografis dalam basis pengetahuan berorientasi tugas yang formal dan dapat dibagikan. Ditunjukkan melalui sistem prototipe, hasilnya memberikan gambaran sekilas yang menjanjikan tentang bagaimana pengguna dapat mulai membangun model pemecahan masalah geografis dan alur kerja yang serupa dengan model dan alur kerja spasial. Model dan alur kerja tersebut dapat digunakan kembali dan diadaptasi untuk permasalahan serupa atau digunakan sebagai landasan untuk mengatasi permasalahan geografis yang lebih kompleks di masa depan, seperti dampak global yang disebabkan oleh perubahan iklim.

Wang dkk. membangun grafik pengetahuan yang mirip dengan Zhuang et al. namun berfokus pada menangkap objek geografis dan konteks spatiotemporalnya. Buku ini menciptakan grafik pengetahuan geografis (GeoKG) yang terdiri dari enam elemen untuk menjawab pertanyaan mendasar dalam geografi termasuk: Dimana letaknya? Mengapa itu ada di sana? Kapan dan bagaimana hal itu terjadi? Melalui proses konstruksi dan formalisasi model, buku ini menangkap objek geografis, relasinya, dan dinamika yang berkelanjutan dalam GeoKG. Untuk menunjukkan keefektifan GeoKG, karya ini merinci evolusi pembagian administratif Nanjing, Tiongkok, di sepanjang Sungai Yangzi dan kemudian membandingkannya dengan ontologi lugas dan dapat diperluas yang dikenal sebagai YAGO (Yet Another Great Ontology). Hasilnya menunjukkan bahwa GeoKG meningkatkan akurasi dan kelengkapan melalui analisis dan evaluasi pengguna, menunjukkan kemajuan ilmiah dalam menangkap pengetahuan geografis dalam sistem komputasi.

1.5 PENCARIAN DATA BESAR

Terakhir, Gaigalas dkk. mempresentasikan pendekatan katalogisasi berbasis infrastruktur siber yang menggabungkan layanan web dan teknologi perayap untuk mendukung pencarian data iklim besar yang efisien. Pendekatan katalogisasi terdiri dari empat langkah utama, termasuk pemilihan dan analisis repositori metadata, perayapan metadata menggunakan crawler, membangun pengindeksan metadata spatiotemporal, dan pencarian berdasarkan pencarian koleksi (melalui layanan katalog) dan pencarian granula (melalui REST API). Pendekatan katalogisasi ini diterapkan untuk mendukung EarthCube

CyberConnector. Untuk menunjukkan kelayakan dan efisiensi pendekatan yang diusulkan, infrastruktur siber ini diuji dengan data ESOM (*Earth System Observation and Modeling*) tingkat petabyte yang disediakan oleh UCAR THREDDS Data Server (TDS). Hasilnya menunjukkan bahwa pendekatan katalogisasi yang diusulkan tidak hanya meningkatkan kecepatan perayapan sebesar 10 kali lipat tetapi juga secara dramatis mengurangi metadata yang berlebihan dari 1,85 gigabyte menjadi 2,2 megabita. Alih-alih berfokus pada analisis big data, penelitian ini menunjukkan signifikansi dan teknik canggih dalam menjadikan big data iklim dapat dicari untuk mendukung kolaborasi antar disiplin ilmu dalam analisis iklim.

1.6 RINGKASAN

Buku ini menyoroti keragaman model dan analisis geografis, data geografis, pemikiran geografis, dan pemikiran komputasi yang digunakan untuk mengatasi berbagai masalah geografis mulai dari mobilitas manusia hingga manajemen bencana. Rancangan buku ini mencakup pemecahan masalah dan pengetahuan geografis, menangani data geografis yang sangat besar dan menganalisis serta memvisualisasikan data geografis.

Kemajuan penting yang disoroti dalam buku ini adalah integrasi pemikiran komputasi dan pemikiran spasial serta penerjemahan ide dan model abstrak ke dalam struktur data dan algoritma yang konkret. Arah penelitian masa depan yang menjanjikan adalah membangun integrasi pengetahuan dan keterampilan lintas disiplin ilmu GIS dan ilmu komputasi, yang disebut literasi siber untuk Ilmu GIS. Dengan cara ini, pengetahuan terintegrasi tentang pola geografis dunia nyata dan proses komputasi dapat ditangkap dan dibagikan, dan data besar serta kerangka analitik visual geografis dapat diintegrasikan untuk menyediakan platform geografis komputasi yang lebih kuat untuk mengatasi berbagai masalah geografis. Tantangan utama dalam arah penelitian ini adalah struktur integratif yang dapat menggabungkan pemikiran ilmiah dengan infrastruktur komputasi, elemen data geografis dengan kemampuan big data, dan metode geografis yang dipadukan dengan paralelisme.

Paralelisme dapat dicapai dengan secara inovatif memanfaatkan kerangka komputasi canggih, seperti MapReduce dan Spark, untuk aplikasi yang mencakup penyortiran data besar-besaran, komputasi, pembelajaran mesin, dan pemrosesan grafik. Meskipun edisi khusus ini menyoroti kemajuan dalam prapemrosesan data besar geografis, prediksi perubahan penggunaan lahan, dan analisis overlay, lebih banyak upaya harus dicurahkan untuk mengidentifikasi aplikasi geografis yang berdampak besar dan memperoleh manfaat dari integrasi geografis. metode dan paralelisasi di era big data. Selain itu, banyak aplikasi data besar geografis yang ada hanya memasukkan tipe atau fungsi data spasial ke dalam sistem data besar yang ada (misalnya Hadoop) tanpa banyak optimasi. Oleh karena itu, arah penelitian lebih lanjut harus fokus pada peningkatan dan optimalisasi kinerja kerangka data besar dari berbagai aspek, seperti ETL data, penjadwalan pekerjaan, alokasi sumber daya, analisis kueri, masalah memori, dan kemacetan I/O, dengan mempertimbangkan prinsip spasial dan kendala.

BAB 2
KERANGKA D_ELT BERBASIS MAPREDUCE DALAM BIG DATA GEOGRAFIS

Sistem *Extracting–Loading–Transforming* (ETL) konvensional biasanya dioperasikan pada satu mesin yang tidak mampu menangani data besar geografis dalam jumlah besar. Untuk menangani sejumlah besar data besar dalam proses ETL, kami mengusulkan D_ELT (*Delayed Extracting–Loading–Transforming*) dengan memanfaatkan paralelisasi berbasis MapReduce. Di antara berbagai jenis big data, kami berkonsentrasi pada big data geografis yang dihasilkan melalui sensor menggunakan teknologi *Internet of Things* (IoT). Dalam lingkungan IoT, latensi pembaruan untuk sensor big data biasanya pendek dan data lama tidak layak untuk dianalisis lebih lanjut, sehingga kecepatan persiapan data menjadi lebih signifikan. Kami melakukan beberapa eksperimen yang mengukur kinerja D_ELT secara keseluruhan dan membandingkannya dengan sistem ETL tradisional dan sistem ekstraksi–pemuatan–transformasi (ELT), menggunakan ukuran data dan tingkat kompleksitas yang berbeda untuk analisis. Hasil eksperimen menunjukkan bahwa D_ELT mengungguli dua pendekatan lainnya, ETL dan ELT. Selain itu, semakin besar jumlah data atau semakin tinggi kompleksitas analisis, semakin besar pula efek paralelisasi transformasi di D_ELT, sehingga menghasilkan kinerja yang lebih baik dibandingkan pendekatan ETL dan ELT tradisional.

2.1 PENDAHULUAN

Dalam beberapa tahun terakhir, berbagai jenis sensor telah terhubung ke Internet of Things (IoT) dan telah menghasilkan data dalam jumlah besar dengan kecepatan tinggi. Sebagian besar data besar sensor ini adalah data geografis, yang menggambarkan informasi tentang benda fisik dalam kaitannya dengan ruang geografis yang dapat direpresentasikan dalam sistem koordinat. Dengan kemajuan teknologi IoT, data yang lebih beragam kini tersedia, sehingga meningkatkan jumlah data besar geografis secara signifikan.

Mengingat sifat umum dari big data, karakteristik unik dari data geografis menciptakan tantangan inovatif dalam persiapan data. Data geografis biasanya mencakup data posisi. Data koordinat ini berbeda dari data string atau integer normal, sehingga memerlukan proses pra-pemrosesan data yang menyertakan banyak perhitungan aritmatika floating-point. Contohnya termasuk transformasi dalam geometri, konversi sistem referensi koordinasi, dan evaluasi hubungan spasial. Diantaranya, aspek data geografis yang paling terkenal adalah hubungan spasial, yang menggambarkan hubungan beberapa objek di lokasi tertentu dengan objek lain di lokasi tetangga. Perhitungan hubungan spasial sebagian besar termasuk dalam analisis spasial dan secara umum dianggap sebagai masalah yang rumit. Selain itu, pengolahan elemen temporal juga mempersulit penanganan data geografis. Untuk menghadapi tantangan dalam pengolahan dan analisis big data geografis, beberapa sistem telah bermunculan. Sistem yang dirancang untuk data besar telah ada selama bertahun-tahun namun, mereka tidak mendapat

informasi tentang properti spasial. Hal ini menyebabkan sejumlah sistem geografis dikembangkan, sebagian besar dengan memasukkan tipe atau fungsi data spasial ke dalam sistem big data yang ada. Hadoop, khususnya, telah terbukti menjadi platform data besar yang matang sehingga beberapa sistem data besar geografis telah dibangun dengan memasukkan kesadaran data spasial ke dalam Hadoop. Namun, masih tidak mudah bagi pengembang software big data untuk membuat aplikasi geografis. Biasanya, untuk menghasilkan pekerjaan MapReduce untuk operasi yang diperlukan di Hadoop, pengembang perlu memprogram fungsi peta dan pengurangan. Analisis spasial biasanya memerlukan penanganan lebih dari satu langkah MapReduce, dimana keluaran data dari langkah MapReduce sebelumnya menjadi masukan ke langkah MapReduce berikutnya. Seiring dengan meningkatnya tingkat kompleksitas analisis spasial, jumlah langkah MapReduce juga meningkat, sehingga menambah kesulitan bagi pengembang untuk menulis kode berulang untuk mendefinisikan langkah-langkah MapReduce yang semakin rumit.

Untuk mengatasi masalah ini, kami menemukan cara untuk merepresentasikan analisis spasial sebagai rangkaian dari satu atau lebih unit operator spasial atau non-spasial. Hal ini memungkinkan pengembang aplikasi big data geografis untuk membuat aplikasi spasial hanya dengan menggabungkan operator spasial atau non-spasial bawaan, tanpa memiliki pengetahuan rinci tentang MapReduce. Setelah rangkaian operator dimasukkan, secara otomatis diubah menjadi peta dan mengurangi pekerjaan di sistem data besar geografis kami yang berbasis Hadoop. Selama proses konversi ini, sistem kami mengontrol jumlah langkah MapReduce sedemikian rupa untuk mencapai kinerja yang lebih baik dengan mengurangi overhead pemetaan dan pengurangan. Tantangan bagi big data geografis, bagaimanapun, terletak pada tidak hanya bagaimana cara menyimpan dan menganalisis data, namun juga bagaimana mentransformasikan data sekaligus mencapai kinerja yang baik.

Saat ini, sejumlah besar data geografis terus disediakan dari banyak sensor spasial. Penting untuk menganalisis data besar geografis ini sesegera mungkin untuk mendapatkan wawasan yang berguna. Namun, waktu yang diperlukan untuk mengubah sejumlah besar data geografis ke dalam platform Hadoop secara bertahap meningkat. Artinya, diperlukan banyak waktu untuk mempersiapkan data-data yang diperlukan untuk analisis geografis, sehingga memperlambat diperolehnya hasil analisis spasial. Misalnya, kami menemukan bahwa diperlukan waktu sekitar 13 jam 30 menit untuk memuat data takograf digital (DTG) sebesar 821 GB menggunakan metode ETL tradisional. Dalam proses ETL, data diekstraksi dari sumber data, kemudian diubah, melibatkan normalisasi dan pembersihan, dan dimuat ke dalam basis data target. Sistem ETL konvensional biasanya dioperasikan pada satu mesin yang tidak dapat secara efektif menangani data besar dalam jumlah besar. Untuk menangani sejumlah besar data besar dalam proses ETL, ada beberapa upaya dalam beberapa tahun terakhir untuk memanfaatkan konsep pemrosesan data paralel.

Usulan ETLMR menggunakan kerangka MapReduce untuk memparalelkan proses ETL. ETLMR dirancang dengan mengintegrasikan MapReduce berbasis Python. Studi ini melakukan evaluasi eksperimental yang menilai skalabilitas sistem berdasarkan skala pekerjaan dan data yang berbeda untuk dibandingkan dengan alat berbasis MapReduce lainnya. Usulan lain

membandingkan solusi ETL berbasis Hadoop dengan solusi ETL komersial dalam hal biaya dan kinerja. Mereka menyimpulkan bahwa solusi ETL berbasis Hadoop lebih baik dibandingkan dengan solusi ETL komersial yang ada. Penelitian lain juga mengimplementasikan P-ETL (parallel-ETL), yang dikembangkan di Hadoop. Dibandingkan dengan tiga langkah tradisional yaitu mengekstraksi, mentransformasikan, dan memuat, P-ETL melibatkan lima langkah yaitu mengekstraksi, mempartisi, mentransformasikan, mereduksi, dan memuat. Studi ini menunjukkan bahwa P-ETL mengungguli skema ETL klasik. Namun, banyak penelitian yang berfokus pada analisis big data, namun penelitian yang berupaya meningkatkan kecepatan penyiapan data yang diperlukan untuk analisis big data masih kurang.

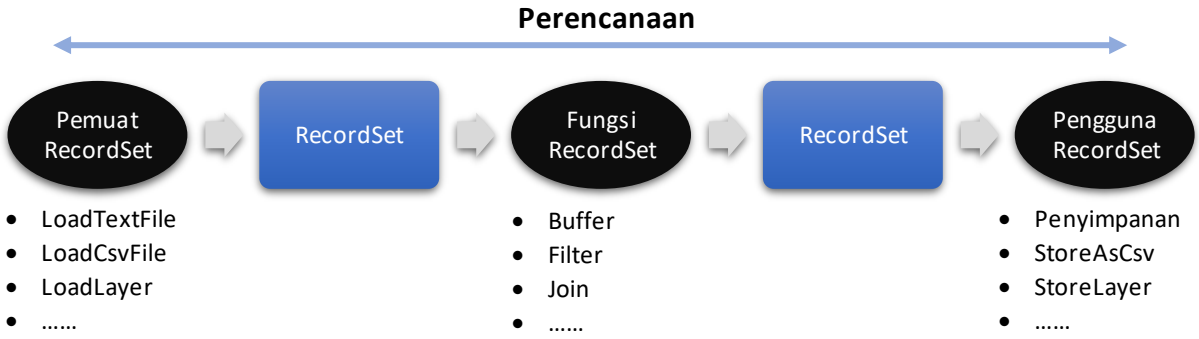
Dalam buku ini, penulis melanjutkan penelitian kami sebelumnya tentang penyimpanan dan pengelolaan data besar geografis dan menjelaskan pendekatan kami untuk meningkatkan kinerja proses ETL. Secara khusus, kami mengusulkan metode untuk memulai analisis data besar geografis dalam waktu singkat dengan mengurangi waktu yang diperlukan untuk transformasi data di bawah platform Hadoop. Transformasi didefinisikan sebagai pemrosesan data yang dicapai dengan mengubah data sumber menjadi format penyimpanan konsisten yang bertujuan untuk melakukan kueri dan menganalisis. Karena sifat transformasi yang kompleks, kinerja proses ETL sangat bergantung pada seberapa efisien transformasi tersebut dilakukan, yang merupakan langkah pembatas laju dalam proses ETL. Pendekatan kami memungkinkan paralelisasi transformasi berbasis MapReduce dalam proses ETL. Di antara berbagai sumber data besar geografis, kami berkonsentrasi pada data besar sensor. Dengan meningkatnya jumlah perangkat penginderaan IoT, jumlah data sensor diperkirakan akan tumbuh secara signifikan dari waktu ke waktu untuk berbagai bidang dan aplikasi. Namun, data sensor berbasis IoT pada dasarnya terstruktur secara longgar dan biasanya tidak lengkap, dan sebagian besar tidak dapat digunakan secara langsung. Selain itu, di lingkungan IoT, periode pembaruan—waktu antara kedatangan data mentah dan saat data penting tersedia—terjadi lebih sering dibandingkan data batch biasa. Kesulitan-kesulitan ini memerlukan penggunaan sumber daya yang besar untuk transformasi dalam proses ETL.

Dalam kajian buku ini memperluas penelitian kami yang disajikan dan menyarankan cara untuk meningkatkan kinerja fungsi transformasi dalam proses ETL dengan memanfaatkan kerangka kerja MapReduce. Pertama, pada bab ini kami menjelaskan secara singkat pekerjaan kami sebelumnya dalam membangun sistem pemrosesan data besar geografis dengan memperluas Hadoop asli untuk mendukung properti spasial. Kami fokus secara khusus pada penjelasan konversi otomatis rangkaian operator yang ditentukan pengguna untuk analisis spasial ke langkah-langkah MapReduce. Dalam bab ini juga menjelaskan penelitian ETL terkini yang diikuti dengan pendekatan kami dalam meningkatkan kinerja transformasi dalam proses ETL berdasarkan MapReduce.

2.2 PLATFORM BIG DATA GEOGRAFIS

Pengembangan sistem pemrosesan data besar geografis berkinerja tinggi berdasarkan Hadoop/MapReduce, bernama Marmot. Di Marmot, analisis spasial didefinisikan sebagai rangkaian RecordSetOperators, di mana RecordSet adalah kumpulan catatan dan

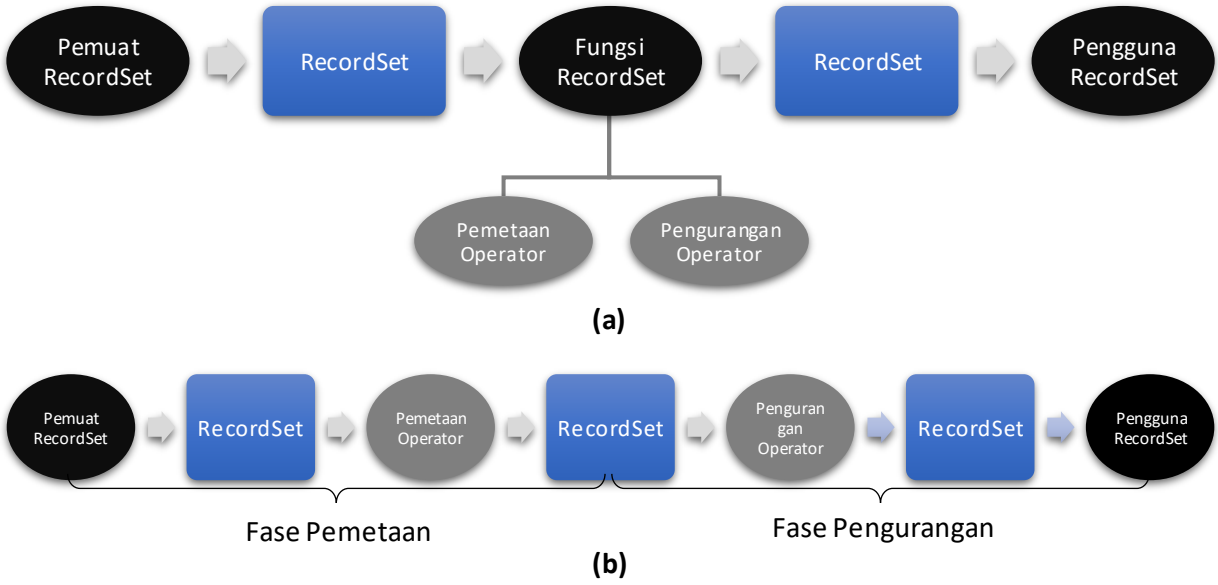
RecordSetOperator adalah elemen pemrosesan menggunakan RecordSet, mirip dengan operator relasional dalam Sistem Manajemen Basis Data Relasional (RDBMS). Urutan RecordSetOperators didefinisikan sebagai Rencana, seperti yang ditunjukkan pada Gambar 2.1.



Gambar 2.1. Representasi analisis spasial di Marmot: Urutan satu atau lebih unit operator spasial atau non-spasial.

Di Marmot, RecordSetOperator diklasifikasikan menjadi tiga kemungkinan tipe: RecordSetLoader, RecordSetFunction, atau RecordSetConsumer. RecordSetLoader adalah operator non-spasial yang memuat data sumber dan mengubahnya menjadi RecordSet; RecordSetFunction adalah operator spasial atau non-spasial yang mengambil RecordSet sebagai data sumber dan menghasilkan RecordSet baru sebagai data keluaran; RecordSetConsumer adalah operator non-spasial yang menyimpan RecordSet yang akhirnya dibuat sebagai hasil analisis spasial tertentu di luar Marmot.

Untuk memproses analisis spasial tertentu, pengembang membuat Rencana terkait dengan menggabungkan operator spasial dan operator non-spasial dan memasukkan Rencana tersebut ke Marmot. Marmot memproses setiap RecordSetOperator satu per satu dan secara otomatis mengubah Rencana yang diberikan untuk memetakan dan mengurangi pekerjaan, seperti yang ditunjukkan pada Gambar 2.2.



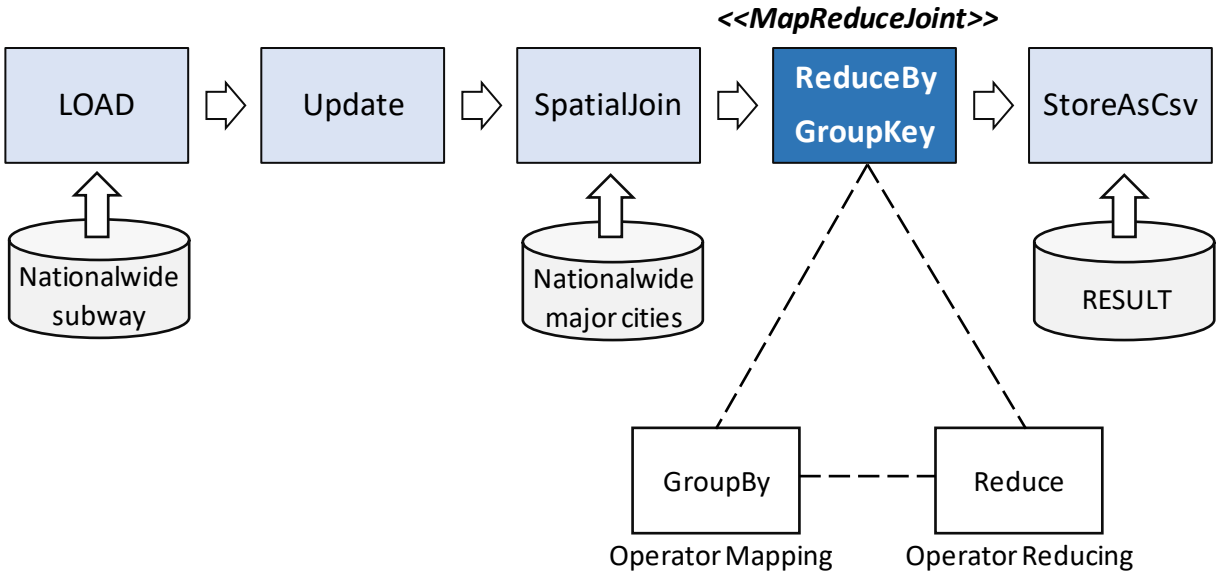
Gambar 2.2. Transformasi otomatis Rencana menjadi pekerjaan MapReduce. (a) Suatu Rencana yang mempunyai RecordSetFunction yang dibagi menjadi operator pemetaan dan pengurangan; (b) Rencana yang diubah secara otomatis.

Saat mengurai Rencana tertentu, ketika Marmot bertemu dengan RecordSetFunction yang dapat dipisahkan menjadi operator pemetaan dan pengurangan (misalnya, ReduceByGroupKey), seperti yang ditunjukkan pada Gambar 2.2a, Marmot menguraikan RecordSetFunction menjadi operator pemetaan dan operator pengurangan, dan pada akhirnya mengubah Rencana menjadi Pekerjaan MapReduce yang terdiri dari fase map dan pengurangan, seperti yang ditunjukkan pada Gambar 2.2b. Selama transformasi ini, Marmot mengontrol jumlah fase MapReduce sedemikian rupa untuk mencapai kinerja yang lebih baik dengan mengurangi overhead pemetaan dan pengurangan. Untuk menjelaskan bagaimana Marmot menangani proses tersebut secara rinci, contoh analisis spasial untuk mengambil stasiun kereta bawah tanah di suatu kota ditunjukkan pada Gambar 2.3 dan 2.4.

```
Plan plan;
plan = marmot.planBuilder("Subways Station per City")
            .load("logs/subways stations")
            .update("the_geom=ST_Centroid(the_geom)")
                  .spatialJoin("the_geom","region/cadastral","the_geom", INTERSECT,
                  "*,param.sig_cd")
            .reduceByGroupKey("sig_cd")
            .aggregate(COUNT())
            .storeAsCsv("result")
            .build();
```

Gambar 2.3. Contoh kode pencarian stasiun kereta bawah tanah per kota.

Gambar 2.3 merupakan kode Marmot sebagai contoh analisis spasial. Analisis direpresentasikan sebagai Rencana yang terdiri dari lima RecordSetOperators: Load, Update, SpatialJoin, ReduceByGroupKey, dan StoreAsCsv. Seperti yang ditunjukkan pada Gambar 2.4, dengan menggunakan operator Beban, Marmot membaca batas setiap stasiun kereta bawah tanah dan menghitung koordinat pusatnya. Titik pusat yang dihitung kemudian digunakan sebagai lokasi perwakilan setiap stasiun kereta bawah tanah melalui operator Update. Untuk setiap stasiun kereta bawah tanah, menggunakan operator SpatialJoin, Marmot mengidentifikasi kota yang menjadi titik pusat stasiun kereta bawah tanah. Terakhir, jumlah stasiun kereta bawah tanah per kota dihitung melalui operator ReduceByGroupKey dan hasilnya disimpan dalam file CSV bernama "result" melalui operator StoreAsCsv.



Gambar 2.4. Contoh Rencana pencarian stasiun kereta bawah tanah per kota.

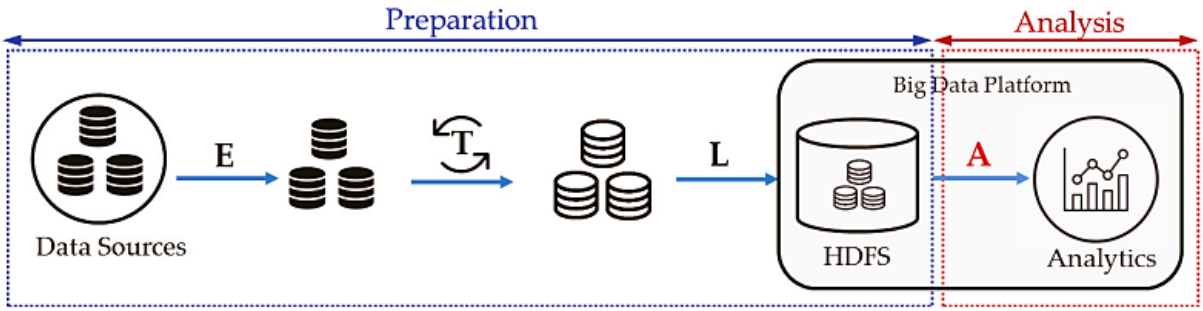
Selama proses transformasi Rencana menjadi rangkaian pekerjaan MapReduce, ReduceByGroupKey didekomposisi menjadi GroupBy dan Reduce masing-masing sebagai operator pemetaan dan operator pereduksi. Oleh karena itu, Load, Update, SpatialJoin, dan GroupBy dijalankan selama fase Map; Kurangi dan StoreAsCsv, selama fase Kurangi.

2.3 KERANGKA D_ELT BERBASIS MAPREDUCE

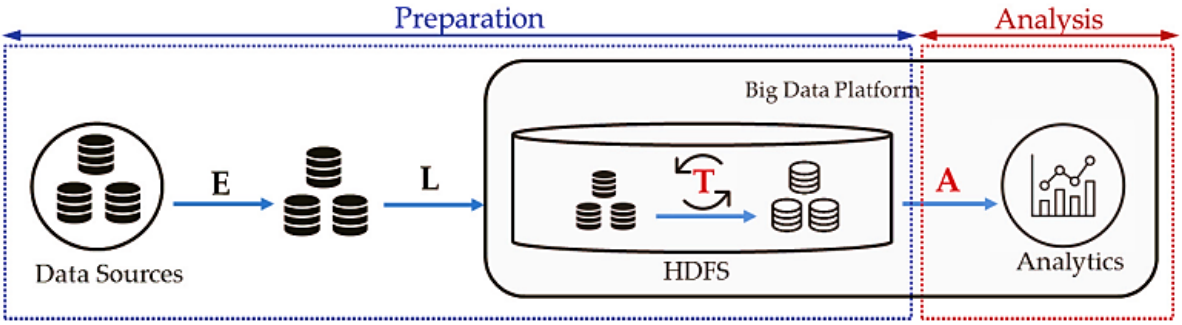
Seperti disebutkan di bagian sebelumnya, penulis membangun Marmot, sistem manajemen data berkinerja tinggi yang memungkinkan pengembang yang tidak memiliki pengetahuan khusus tentang teknologi big data untuk mengimplementasikan aplikasi analisis spasial berkinerja tinggi ke big data geografis. Namun permasalahan terkait big data geografis tidak hanya terletak pada cara mengelola data secara efisien untuk analisis cepat, namun juga pada cara mentransformasikan data secara efisien untuk persiapan data yang cepat.

Data DTG, misalnya, telah digunakan untuk menganalisis status operasional transportasi untuk mengidentifikasi titik-titik perbaikan dan mengidentifikasi daerah-daerah tertinggal dalam hal transportasi umum. Otoritas transportasi, misalnya Otoritas Keselamatan Transportasi Korea, mengumpulkan data DTG dari kendaraan komersial dan menerapkan analisis pada data besar tersebut untuk mengekstraksi wawasan dan memfasilitasi pengambilan keputusan. Seringkali, hasil analisis data harus diperoleh secara berkala dalam waktu tertentu, misalnya setiap hari, agar siap menghadapi kasus-kasus yang muncul. Dalam situasi ini, untuk menyelesaikan analisis yang diberikan tepat waktu, tidak hanya kecepatan analisis data, namun juga kecepatan persiapan data merupakan faktor penting yang mempengaruhi kinerja secara keseluruhan. Dalam lingkungan IoT, latensi pembaruan untuk data besar sensor, yang menjadi fokus buku ini ini di antara berbagai sumber data besar geografis, biasanya berupa data pendek dan lama yang tidak layak untuk dianalisis lebih lanjut, sehingga kecepatan persiapan data menjadi lebih penting. Selain itu, data besar sensor

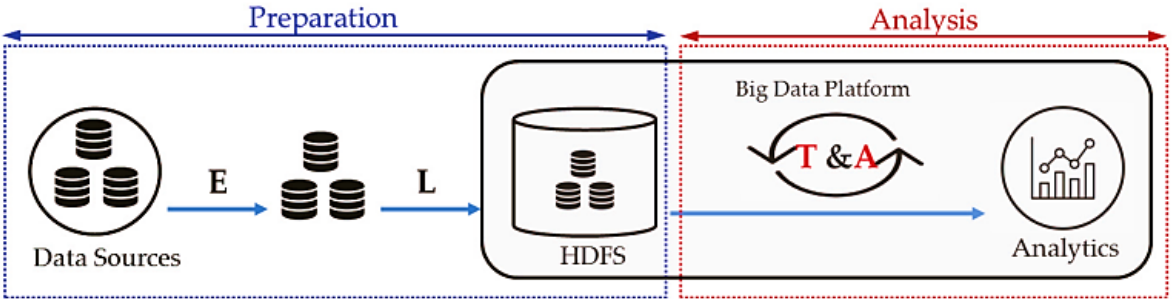
dihasilkan oleh mesin; oleh karena itu, data sumber mengandung lebih banyak gangguan atau kesalahan dibandingkan dengan data yang dihasilkan manusia, sehingga semakin mempersulit persiapan data.



a) Proses ETL



b) Proses ELT



c) Proses D_ELT

Gambar 2.5. Ilustrasi proses persiapan dan analisis big data geografis yang membandingkan tiga kasus: (a) ETL; (b) ELT; (c) D_ELT. Pada gambar, "E" berarti ekstrak, "T" berarti transformasi, "L" berarti beban, dan "A" berarti analisis.

ETL tradisional tidak lagi dapat mengakomodasi situasi seperti itu. ETL dirancang untuk komputasi ringan pada kumpulan data kecil, namun tidak mampu menangani data dalam jumlah besar secara efisien. Gambar 2.5a menjelaskan persiapan dan analisis data dalam proses ETL. Dalam pendekatan ini, data diekstraksi dari berbagai sumber dan kemudian diubah pada server ETL, yang biasanya berupa satu mesin, dan dimuat ke dalam sistem file terdistribusi Hadoop (HDFS). Data yang dimuat akhirnya dianalisis dalam platform data besar untuk pengambilan keputusan. Dalam pendekatan ini, operasi analisis diproses secara paralel/terdistribusi menggunakan MapReduce, yang menjamin kinerja yang wajar, namun kemacetan dapat terjadi selama operasi transformasi. Faktanya, transformasi adalah fase

yang paling memakan waktu dalam ETL karena operasi ini mencakup penyaringan atau agregasi data sumber agar sesuai dengan struktur database target. Pembersihan data juga harus diselesaikan untuk setiap data duplikat, data yang hilang, atau format data yang berbeda. Terlebih lagi, dalam lingkungan big data, karena sumber big data yang heterogen, operasi transformasi tradisional akan menciptakan lebih banyak beban komputasi. Oleh karena itu, kinerja keseluruhan proses ETL terutama bergantung pada seberapa efisien operasi transformasi dilakukan.

Untuk mengatasi kelemahan ETL tradisional dan untuk mempercepat proses persiapan data, proses ELT dirancang. Sifat ETL tradisional adalah melakukan transformasi segera setelah operasi ekstrak dan kemudian memulai operasi pemuatan. Sebaliknya, ide dasar ELT adalah melakukan operasi pemuatan segera setelah operasi ekstrak, dan melakukan transformasi setelah menyimpan data di HDFS, seperti yang ditunjukkan pada Gambar 2.5b. Pendekatan ini memiliki beberapa keunggulan dibandingkan ETL. Operasi transformasi dapat dilakukan pada saat run time bila diperlukan dan dimungkinkan untuk menggunakan transformasi bahkan beberapa kali untuk menangani perubahan kebutuhan data. Selain itu, pendekatan ini menghilangkan mesin transformasi terpisah, server ETL, antara sumber dan target dan menjadikan keseluruhan sistem lebih murah. Yang terpenting, ELT memungkinkan data sumber mentah dimuat langsung ke target dan juga memanfaatkan sistem target untuk melakukan operasi transformasi. Dalam hal ini, ELT dapat mempercepat transformasi menggunakan paralelisasi/distribusi yang didukung platform big data berbasis Hadoop.

Terlepas dari kelebihan tersebut, ELT masih memiliki keterbatasan dalam menangani big data. Kerangka kerja ELT dapat mempercepat transformasi menggunakan MapReduce, namun analisis dimulai hanya setelah transformasi selesai. Dalam pendekatan ini, sulit untuk mengoptimalkan transformasi bersamaan dengan analisis karena transformasi dilakukan secara batch, apa pun konteks analisisnya. Misalnya, dalam kasus data geografis, salah satu overhead komputasi yang tinggi dalam melakukan transformasi terjadi selama transformasi tipe, seperti mengubah sumbu x–dan y–teks biasa menjadi (x,y) koordinat titik dan transformasi sistem koordinat untuk melakukan analisis spasial. Jika analisis tidak memerlukan tugas seperti itu, tugas tersebut dapat diidentifikasi pada fase transformasi dan hanya memuat data yang diperlukan. Dengan melakukan hal ini, sistem dapat menghilangkan transformasi yang tidak perlu dan mempercepat kinerja.

Untuk mencapai skalabilitas dan kinerja yang lebih baik dalam melakukan transformasi pada big data geografis, buku ini ini menawarkan pendekatan baru untuk persiapan data yang disebut D_ETL—dalam arti bahwa keputusan tentang bagaimana melakukan transformasi ditunda hingga konteks analisisnya dipahami. Seperti yang ditunjukkan pada Gambar 2.5c, dalam pendekatan kami, transformasi dijalankan secara paralel/terdistribusi dengan analisis dalam platform data besar geografis kami, Marmot. Di Marmot, operator untuk transformasi dianggap sebagai jenis RecordSetOperator dan juga terdiri dari Rencana, bersama dengan RecordSetOperator yang dirancang untuk analisis. Pendekatan ini memiliki keuntungan karena proses persiapan dan analisis data dijelaskan menggunakan model data yang sama.

Oleh karena itu, pengembang aplikasi dapat terbebas dari ketidaknyamanan karena harus terbiasa mengimplementasikan kedua proses tersebut.

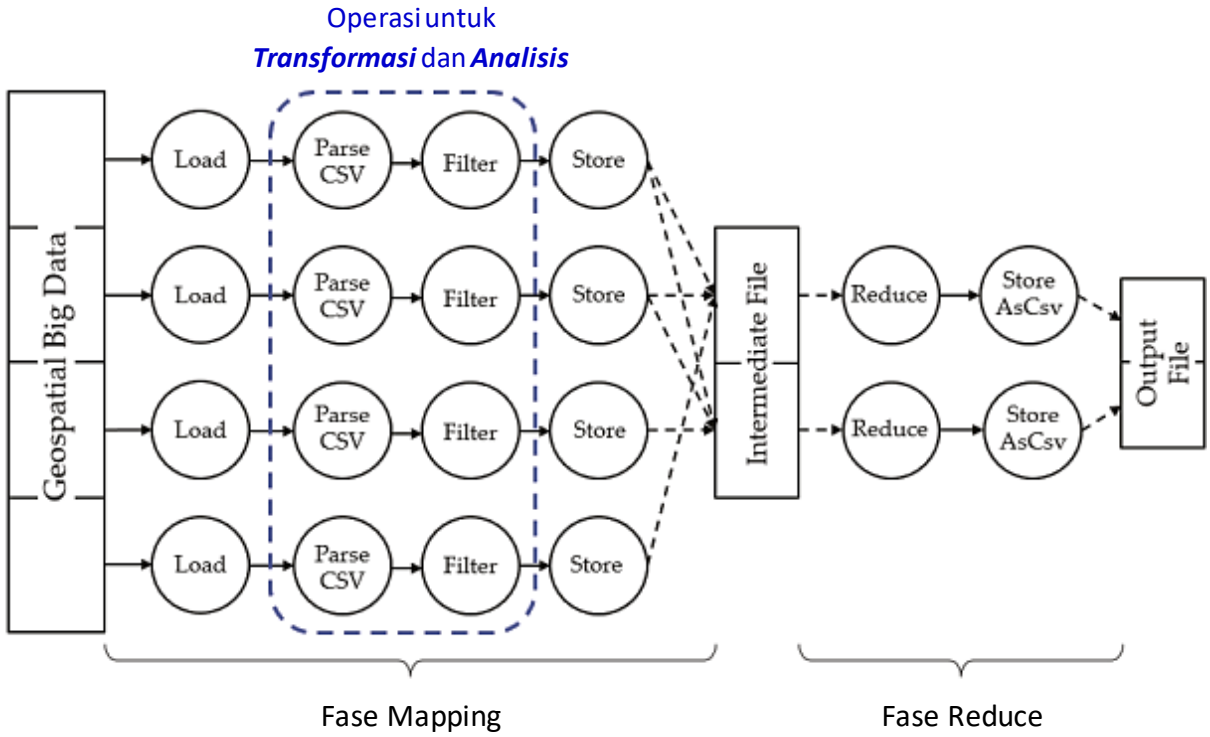
Mengenai operator yang diperlukan untuk melakukan transformasi, pengembang aplikasi menentukannya dalam skrip D_ELT. Dengan cara ini, pengembang dapat mengimplementasikan persiapan dan analisis data secara bersamaan, tanpa harus mengubah kode yang ada untuk melakukan analisis. Skrip D_ELT terdiri dari nama operator dan daftar nilai kunci parameter, seperti yang ditunjukkan pada Gambar 2.6. Untuk memudahkan, jika pengembang memerlukan operator baru untuk melakukan transformasi, operator dapat diimplementasikan secara terpisah sebagai formulir plug-in dan dapat digunakan di Marmot, dengan cara yang sama seperti untuk operator yang ada.

```
{
  "name": "import_plan".
  "operator": [{
    "parseCsv": {
      "delimiter": ",",
      "options": {
        "headerColumn": ["car_no","ts", "month", "sid_cd", "besselX", "besselY", "status",
        "company", "driver_id", "xpos", "ypos"],
        "commentMarker": "#"
      }
    }
}, {
  "expand": {
    "column": [{
      "name": "status"
    }]
  }
}, {
  "toPoint": {
    "xColumn": "xpos",
    "yColumn": "ypos",
    "outColumn": "the_geom"
  }
}, {
  "transformCrs": {
    "geometryColumn": "the_geom",
    "sourceSrid": "EPSG:4326",
    "targetSrid": "EPSG:5186"
  }
}, {
  "project": {
    "columnExpr": "the_geom, *-
      {the_geom,xpos,ypos,besselX,besselY,month,sid_cd}"
    }
  } ]
```

}

Gambar 2.6. Contoh skrip D_ELT (delayedextraction–loading–transforming) yang menjelaskan operator yang diperlukan untuk transformasi data.

Untuk melakukan analisis spasial, Marmot terlebih dahulu memuat skrip D_ELT untuk menentukan operator apa yang perlu dijalankan untuk transformasi. Kemudian, Marmot (1) memeriksa operator yang perlu dijalankan untuk analisis, (2) hanya memuat data yang diperlukan berdasarkan kebutuhan analisis, dan (3) menjalankan transformasi dan analisis dengan cara terdistribusi paralel. Pada saat ini, sebagian dari data yang diubah dapat digunakan untuk analisis dan tidak perlu menunggu sampai seluruh data selesai diubah. Gambar 2.7 menunjukkan urutan operator yang dieksekusi untuk transformasi dan analisis serta komposisinya sebagai bentuk rencana. Dalam contoh Rencana ini, "ParseCSV" adalah operator untuk transformasi dan "Filter" adalah operator untuk analisis. Mereka dialokasikan dalam fase Peta dan dieksekusi dengan cara terdistribusi paralel. Keluaran dari fase Peta digabungkan selama fase Pengurangan dan hasilnya ditulis dalam file keluaran.



Gambar 2.7. Ilustrasi tahapan Map dan Reduce pada proses D_ELT.

Alasan mengapa kami menerapkan D_ELT menggunakan MapReduce dan bukan Spark, mesin terkenal lainnya untuk pemrosesan data besar, adalah karena platform besar geografis yang kami kembangkan sebelumnya didasarkan pada Hadoop dan kami memiliki tujuan untuk meningkatkan waktu transformasi data di lingkungan tersebut. Selain itu, data yang kami tangani saat ini adalah data DTG dalam jumlah besar, yang menghasilkan 20–30 TB setiap bulannya. Dengan menggunakan Spark, saat menjalankan analisis spasial berdasarkan data

berukuran besar ini, kami mengantisipasi kemungkinan terjadinya masalah yang tidak terduga (misalnya, pertukaran disk), namun sepengetahuan kami, solusi konkrit belum diusulkan.

Penting juga untuk dicatat bahwa ELT dan D_ELT identik dalam hal melakukan transformasi data selama fase MapReduce di Hadoop. Perbedaan antara ELT dan D_ELT adalah sebagai berikut. Di ELT, setelah data mentah diunggah ke Hadoop, data tersebut diubah menggunakan MapReduce. Setelah transformasi benar-benar selesai, analisis kemudian dimulai menggunakan MapReduce yang lain. Namun pada D_ELT tidak dilakukan transformasi data, meskipun seluruh data mentah diunggah ke Hadoop namun tertunda hingga saat dilakukan analisis. Artinya, tugas transformasi ditumpangi ke tugas analisis dan kedua tugas tersebut dilakukan bersama-sama menggunakan MapReduce yang sama. Dengan cara ini, sebagian data yang diubah dapat segera digunakan untuk analisis tanpa harus menunggu seluruh data diubah.

2.4 EVALUASI EKSPERIMENTAL

Bagian ini menjelaskan evaluasi kami terhadap peningkatan kinerja yang dicapai dengan pendekatan yang kami usulkan, D_ELT. Selain itu, skalabilitas dari tiga pendekatan berbeda (ETL tradisional, ELT, dan D_ELT) diukur dan dibandingkan dengan memvariasikan ukuran data dan tingkat kompleksitas analisis.

Pengaturan eksperimen

Eksperimen kami dilakukan pada empat node cluster Hadoop. Setiap node adalah komputer desktop dengan CPU Intel 4 core i7 4,0 GHZ, memori utama 32 GB, dan disk 4 TB. Sistem operasinya adalah CentOS 6.9 dan versi Hadoop adalah Hortonworks HDP 2.6.1.0 dengan Ambari 2.5.0.3. PostgreSQL 9.5 digunakan untuk sistem manajemen basis data bersama dengan Oracle JDK 1.8. MapReduce2 versi 2.7.3 digunakan.

Data uji yang digunakan dalam percobaan adalah data DTG yang dipasang pada kendaraan, yang mencatat catatan berkendara secara real time. Struktur datanya terdiri dari stempel waktu, nomor kendaraan, jarak tempuh harian, akumulasi jarak tempuh, kecepatan, akselerasi, RPM, rem, posisi_x, posisi_y, dan sudut. Datanya diklasifikasikan ke dalam tiga ukuran berbeda: kecil, 9,9 GB; sedang, 19,8GB; dan besar, 29,8 GB, seperti yang ditunjukkan pada Tabel 2.1. Untuk platform data besar geografis, kami menggunakan sistem Marmot yang kami kembangkan.

Tabel 2.1. Ukuran data: kecil, sedang, dan besar.

	Ukuran Data
Kecil	9,9GB
Sedang	19,8GB
Besar	29,8 GB

Eksperimen 1: Pengukuran Waktu Persiapan Data

Dalam eksperimen ini, kami membandingkan waktu persiapan data ETL dan ELT dengan D_ELT yang kami usulkan, dan skalabilitas setiap pendekatan berdasarkan ukuran data yang berbeda. Hasil keseluruhan dari percobaan ini disajikan pada Tabel 2.2.

Tabel 2.2. Waktu persiapan data (dalam detik): ETL, ELT, dan D_ELT.

	ETL [1]	**ELT [2]**	**H_ELT [3]**
Kecil	579	413	116
Sedang	1158	808	231
Besar	1727	1175	345

Seperti terlihat pada Gambar 2.5, total waktu penyiapan data pada proses ETL meliputi waktu ekstraksi, transformasi, dan pemuatan. Dalam kasus ELT, total waktu yang dihabiskan untuk persiapan data adalah penjumlahan waktu ekstraksi, pemuatan, dan transformasi. Sedangkan pada proses ETL, transformasi dilakukan pada satu mesin yang berbasis non-MapReduce dan transformasi pada ELT dilakukan secara terdistribusi secara paralel berbasis MapReduce.

1. Waktu penyiapan data dalam ETL: E+T+L, dimana E untuk ekstrak, T untuk transformasi, L untuk beban;
2. Waktu persiapan data dalam ELT: E+L+T, di mana T dieksekusi secara terdistribusi paralel;
3. Waktu persiapan data di D_ELT: E+L.

Dalam kasus D_ELT, total waktu yang dihabiskan untuk persiapan data adalah penjumlahan waktu hanya untuk mengekstraksi dan memuat, namun tidak termasuk waktu untuk mentransformasikannya. Transformasi dilakukan bersamaan dengan analisis pada tahap analisis data, sehingga penyiapan data di D_ELT tidak meliputi transformasi melainkan hanya ekstrak dan pemuatan.

Eksperimen 2: Pengukuran Waktu Analisis Data

Dalam percobaan ini, kami membandingkan waktu analisis data ETL, ELT, dan D_ELT yang kami usulkan, serta skalabilitas setiap pendekatan berdasarkan ukuran data yang berbeda. Selain itu, eksperimen ini juga menyertakan D_ELT, D_ELT_Opt yang dioptimalkan, melakukan analisis data dengan memfilter dan hanya menggunakan data yang diperlukan. Untuk melihat variasi performa menurut tingkat kompleksitas analisis yang berbeda, kami menggunakan tiga analisis—Count, GroupBy, dan SpatialJoin—masing-masing untuk analisis kompleks tingkat rendah, menengah, dan tinggi. Hasil keseluruhan dari percobaan ini disajikan pada Tabel 2.3.

Tabel 2.3. Waktu analisis data (dalam detik): ETL(atau ELT), D_ELT, dan D_ELT yang dioptimalkan.

		ETL(atau ELT)[1]	**D_ELT [2]**	**D_ELT_Opt[3]**
Count	Kecil	68	96	57

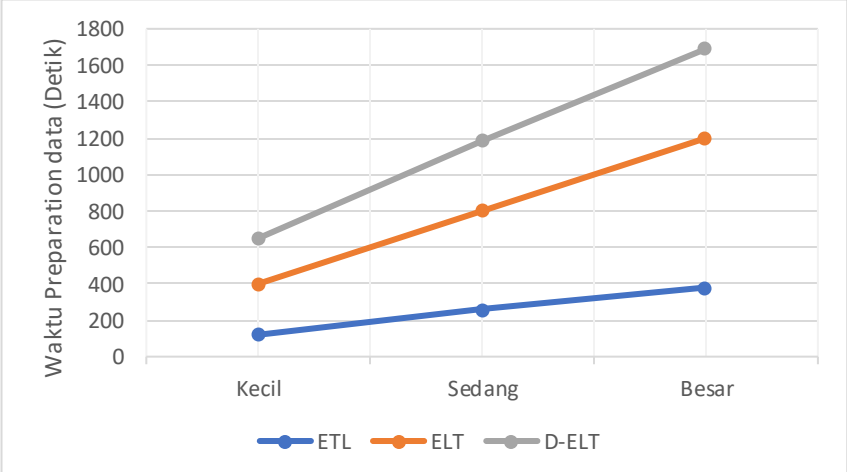
	Sedang	126	179	104
	Besar	181	257	143
GroupBy	Kecil	76	98	87
	Sedang	139	179	162
	Besar	203	256	233
SpatialJoin	Kecil	391	406	406
	Sedang	772	806	806
	Besar	1087	1190	1148

1. Waktu analisis data dalam ETL atau ELT: A, dimana A dijalankan secara paralel;
2. Waktu analisis data di D_ELT: T+A, T, dimana T, A dieksekusi secara paralel;
3. Waktu analisis data dalam D_ELT yang dioptimalkan: T+A, di mana T, A dijalankan secara paralel hanya dengan menggunakan data yang diperlukan.

Seperti yang ditunjukkan pada Gambar 2.5, total waktu untuk analisis data di ETL hanya mencakup analisis, yang dilakukan secara terdistribusi paralel menggunakan MapReduce. Dalam kasus ELT, setelah persiapan data selesai, analisis akan dilakukan dengan cara yang sama seperti ETL. Dalam kasus D_ELT dan D_ELT yang dioptimalkan, total waktu yang dihabiskan untuk analisis data adalah waktu yang diperlukan untuk menjalankan transformasi dan analisis secara terdistribusi paralel menggunakan MapReduce.

2.5 IMPLEMENTASI DAN TEMUAN

Eksperimen pertama untuk mengukur waktu persiapan data untuk setiap pendekatan mengungkapkan hal-hal berikut. Seperti yang ditunjukkan pada Gambar 2.8, D_ELT sekitar 5 kali lebih cepat dibandingkan ETL (116 detik vs. 579 detik untuk data kecil; 231 detik vs.1158 detik untuk data sedang; 345 detik vs. 1727 detik untuk data besar) dan sekitar 3 kali lebih cepat dibandingkan ELT (116 detik vs. 413 detik untuk data kecil; 231 detik vs. 808 detik untuk data sedang; 345 detik vs. 1175 detik untuk data besar), berapa pun ukuran datanya. Pendekatan ELT sekitar 1,4 kali lebih cepat dibandingkan ETL. Hal ini karena efek pemrosesan terdistribusi paralel menggunakan MapReduce Marmot saat melakukan transformasi di ELT.



Gambar 2.8. Waktu persiapan data (dalam detik): ETL, ELT, dan D_ELT.

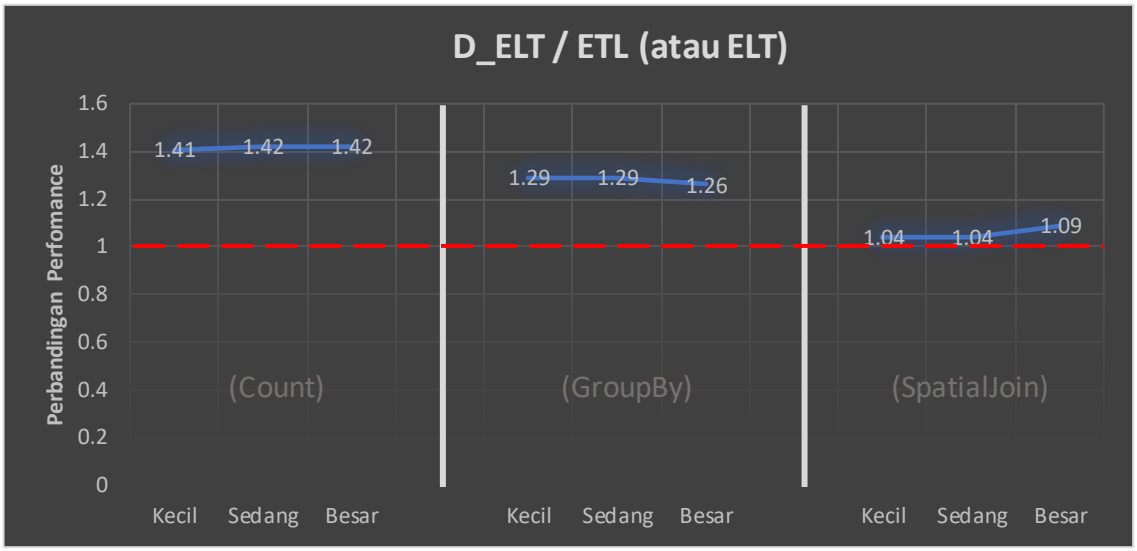
Waktu analisis data diukur dengan percobaan kedua dan mengungkapkan poin-poin berikut. Tabel 2.4 membandingkan kinerja antara D_ELT dan ETL(atau ELT) dan D_ELT dan ETL(atau ELT) yang dioptimalkan. Dalam kedua kasus tersebut, rasio kinerja terhadap analisis hampir sama terlepas dari ukuran datanya. Hal yang menarik adalah waktu analisis data pada proses D_ELT berisi waktu untuk transformasi, sedangkan proses ETL (atau ELT) tidak menyertakan waktu tersebut. Meskipun D_ELT lebih lambat dibandingkan ETL(atau ELT), hanya ada sedikit perbedaan dalam performa—D_ELT hingga 1,4 kali lebih lambat. Dalam kasus D_ELT yang dioptimalkan, prosesnya hanya 1,2 kali lebih lambat dibandingkan pendekatan ETL (atau ELT). Di D_ELT, dalam kasus analisis sederhana, waktu yang dibutuhkan dalam transformasi data relatif besar dibandingkan dengan waktu analisis dan menghabiskan sebagian besar waktu eksekusi total. Namun, dalam kasus analisis spasial yang kompleks, waktu yang dibutuhkan dalam transformasi data relatif kecil dibandingkan dengan analisis data, sehingga biaya transformasi yang dikeluarkan juga relatif kecil.

Tabel 2.4. Perbandingan kinerja: D_ELT/ETL(atau ELT) dan D_ELT/ETL(atau ELT) yang dioptimalkan.

		D_ELT/ETL(atau ELT)	D_ELT_Opt/ETL(atau ELT)
Count	Kecil	1.41	0.84
	Sedang	1.42	0.83
	Besar	1.42	0.79
GroupBy	Kecil	1.29	1.14
	Sedang	1.29	1.17
	Besar	1.26	1.15
SpatialJoin	Kecil	1.04	1.04
	Sedang	1.04	1.04
	Besar	1.09	1.06

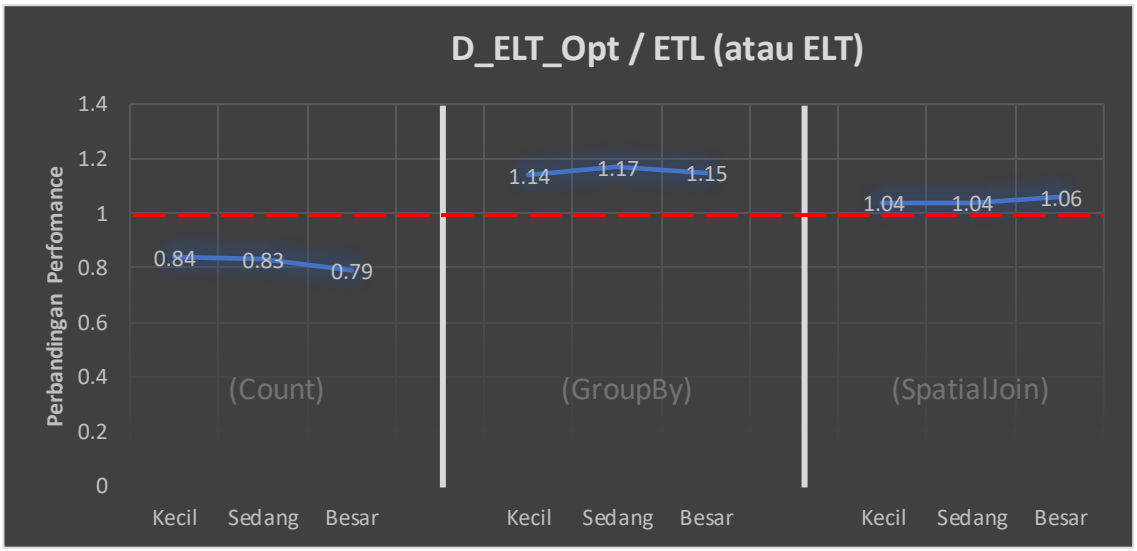
Penting untuk dicatat bahwa hal ini tidak berdampak pada penurunan kinerja secara keseluruhan, mengingat D_ELT sekitar 3–5 kali lebih cepat dibandingkan ETL(atau ELT) selama persiapan data, seperti yang ditunjukkan pada Tabel 2.2 dan Gambar 2.8. Oleh karena itu, kinerja D_ELT dan D_ELT yang dioptimalkan jauh lebih besar dibandingkan ETL atau ELT.

Gambar 2.9 membandingkan kinerja antara D_ELT dan ETL (atau ELT) selama analisis data menurut ukuran data dan jenis analisis yang berbeda. Seperti yang telah disebutkan sebelumnya, waktu analisis dalam D_ELT mencakup transformasi dibandingkan dengan ETL(atau ELT), sehingga D_ELT lebih lambat dibandingkan ETL(atau ELT), seperti yang ditunjukkan pada Tabel 2.3. Namun, semakin tinggi kompleksitas analisisnya (Hitung <GrupBerdasarkan< SpatialJoin), semakin kecil perbedaan antara kinerja D_ELT dan ETL (atau ELT). Alasannya adalah semakin tinggi kompleksitas analisis, semakin tinggi pula efek paralelisasi transformasi D_ELT, sehingga meningkatkan kinerja D_ELT.



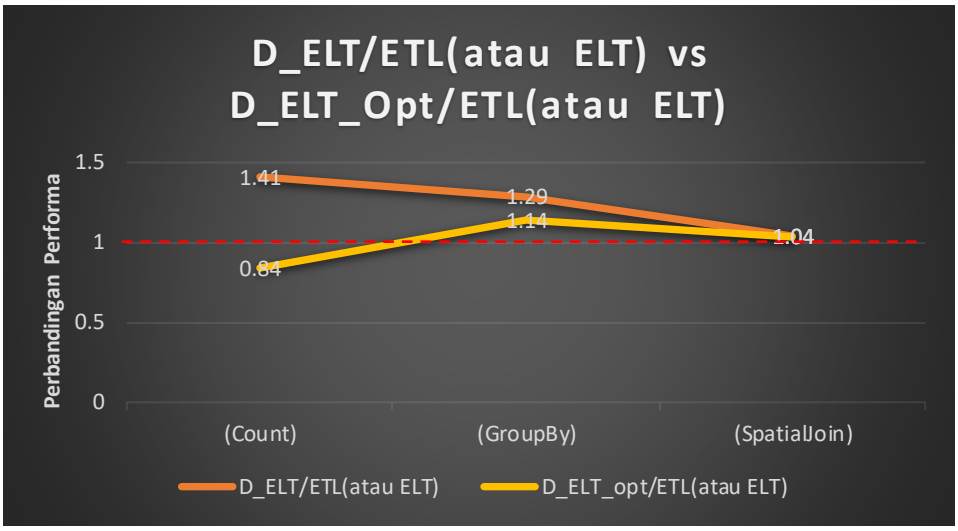
Gambar 2.9. Perbandingan performa D_ELT/ETL(atau ELT) berdasarkan ukuran data kecil, sedang, dan besar untuk masing-masing dari tiga analisis: Hitungan, Kelompok-Berdasarkan, dan SpasialJoin.

Demikian pula, Gambar 2.10 membandingkan kinerja antara D_ELT dan ETL (atau ELT) yang dioptimalkan selama analisis data, menurut ukuran data dan jenis analisis yang berbeda. Dibandingkan dengan Gambar 2.9, dalam kasus dua kasus analisis sederhana, Count dan GroupBy, D_ELT yang dioptimalkan lebih cepat daripada D_ELT. Hal ini karena untuk analisis sederhana, sejumlah besar data sering kali tidak terkait dengan analisis, sehingga lebih banyak data dapat disertakan dalam target pengoptimalan, sehingga menghasilkan peningkatan bertahap dalam performa D_ELT yang dioptimalkan.



Gambar 10. Perbandingan performa D_ELT/ETL(atau ELT) yang dioptimalkan berdasarkan ukuran data kecil, sedang, dan besar untuk masing-masing dari tiga analisis: Hitungan, Kelompok-Berdasarkan, dan SpatialJoin.

Dalam kedua kasus tersebut, rasio kinerja analisisnya sangat mirip, berapa pun ukuran datanya. Oleh karena itu, kami hanya memilih ukuran data yang kecil untuk membandingkan rasio kinerja, seperti yang ditunjukkan pada Gambar 2.11. Hal ini menunjukkan bahwa semakin tinggi kompleksitas analisis, semakin kecil perbedaan kinerja antara D_ELT dan D_ELT yang dioptimalkan. Hal ini karena semakin kompleks analisisnya, semakin banyak data yang terlibat dalam analisis, sehingga mengurangi cakupan pengoptimalan. Dalam kasus SpatialJoin, yang memiliki kompleksitas tertinggi di antara ketiga analisis, kedua nilai pada Gambar 2.11 menyatu hingga hampir 1,0, menunjukkan bahwa hampir tidak ada perbedaan kinerja antara D_ELT dan D_ELT yang dioptimalkan.

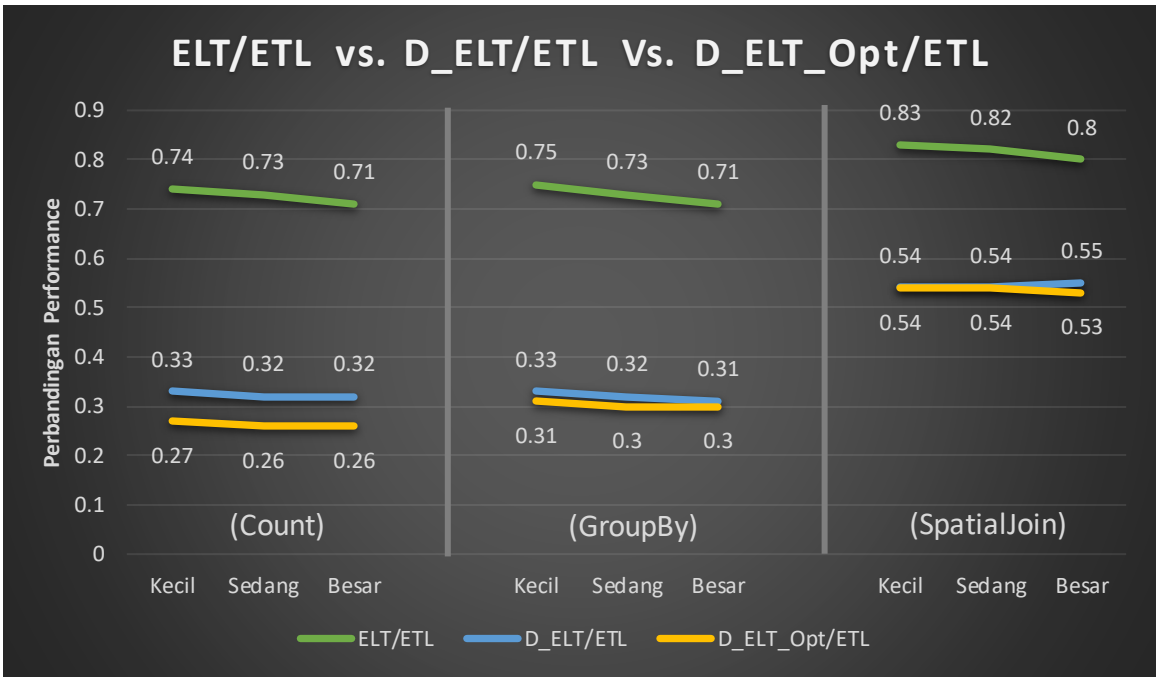


Gambar 11. Perbandingan kinerja D_ELT/ETL(atau ELT) vs. D_ELT/ETL(atau ELT) yang dioptimalkan berdasarkan pada ukuran data kecil untuk masing-masing dari tiga analisis: Hitungan, Kelompok-Berdasarkan, dan SpatialJoin.

Performa keseluruhan ETL, ELT, D_ELT, dan D_ELT yang dioptimalkan diperoleh dengan menjumlahkan waktu persiapan dan analisis data. Tabel 2.5 menunjukkan bahwa kinerja D_ELT secara keseluruhan jauh lebih cepat dibandingkan dengan pendekatan ETL atau ELT. D_ELT hingga 3 kali lebih cepat dari ETL dan 2 kali lebih cepat dari ELT. D_ELT yang dioptimalkan hingga 4 kali lebih cepat dari ETL dan 3 kali lebih cepat dari ELT. Hasilnya diperoleh dari dua kasus analisis sederhana, Count dan GroupBy, tetapi tidak SpatialJoin. Dalam kasus SpatialJoin, baik D_ELT maupun D_ELT yang dioptimalkan masih memiliki performa lebih baik dibandingkan ETL atau ELT, namun hampir tidak ada perbedaan antara performa keseluruhan D_ELT dan D_ELT yang dioptimalkan. Gambar 2.12 menunjukkan bahwa seiring dengan meningkatnya kompleksitas analisis, kesenjangan antara D_ELT dan D_ELT yang dioptimalkan semakin berkurang.

Tabel 2.5. Kinerja keseluruhan ETL, ELT, D_ELT, D_ELT yang dioptimalkan (dalam hitungan detik), dan perbandingan kinerja antara ELT vs. ETL, D_ELT vs. ETL, dan D_ELT vs. ETL yang dioptimalkan.

		ETL	ELT	D_ELT	D_ELT_Opt	ELT/ETL	D_ELT/ETL	D_ELT_Opt/ETL
Count	Kecil	647	481	212	173	0.74	0.33	0.27
	Sedang	1284	934	410	335	0.73	0.32	0.26
	Besar	1908	1356	602	488	0.71	0.32	0.26
GroupBy	Kecil	655	489	214	203	0.75	0.33	0.31
	Sedang	1297	947	410	393	0.73	0.32	0.30
	Besar	1930	1378	601	578	0.71	0.31	0.30
SpatialJoin	Kecil	970	804	522	522	0.83	0.54	0.54
	Sedang	1930	1580	1037	1037	0.82	0.54	0.54
	Besar	2814	2262	1535	1493	0.80	0.55	0.53



Gambar 12. Perbandingan kinerja keseluruhan ELT/ETL vs. D_ELT/ETL vs. D_ELT/ETL yang dioptimalkan berdasarkan ukuran data kecil, sedang, dan besar untuk masing-masing dari tiga analisis: Hitungan, Kelompok-Berdasarkan, dan SpatialJoin.

Ada dua metode konvensional—ETL dan ELT. Metode ETL tradisional tidak menggunakan metode terdistribusi/paralel pada saat pra-pemrosesan data sehingga menimbulkan masalah terutama ketika volume data yang akan dipra-pemrosesan besar. Metode ELT menyempurnakan metode ETL tradisional untuk mempercepat pra-pemrosesan data menggunakan metode terdistribusi/paralel. Metode D_ELT yang kami usulkan mengurangi overhead dalam pra-pemrosesan data. Di D_ELT, tugas transformasi ditumpangkan ke tugas analisis dan kedua tugas tersebut dilakukan bersama-sama menggunakan MapReduce yang sama. Cara ini memungkinkan seseorang untuk melakukan analisis dengan segera tanpa menyimpan hasil transformasi dan juga mengecualikan transformasi yang tidak perlu yang tidak digunakan dalam analisis.

Dibandingkan dengan metode yang ada, metode D_ELT secara signifikan mengurangi waktu persiapan data, namun memiliki kelemahan dalam kasus berikut. Pertama, analisis yang sama harus dilakukan berulang-ulang. Misalnya, metode D_ELT menghasilkan pengurangan waktu persiapan data sebesar 1382 detik (data besar, Tabel 2.2) dibandingkan dengan metode ETL konvensional, namun ditambahkan 103 detik (data besar, SpatialJoin, Tabel 2.3) setiap kali analisis dilakukan. Oleh karena itu, semakin besar jumlah analisis yang dilakukan, D_ELT semakin tidak efisien dibandingkan dengan metode tradisional. Pada contoh di atas, metode D_ELT lebih tidak efisien dibandingkan metode yang sudah ada bila analisis yang sama dilakukan lebih dari 14 kali berturut-turut. Kedua, jika sejumlah besar data masukan tidak valid, sejumlah besar data dapat dihapus sebagai hasil transformasi. Di D_ELT, tugas transformasi dipikul setiap kali tugas analisis dijalankan, sejumlah besar data yang tidak valid dibaca berulang kali, sehingga mengakibatkan beban I/O dan komputasi yang tidak diperlukan. Akhirnya, metode yang diusulkan dalam buku ini ini tidak mempertimbangkan aplikasi real-time. Namun, metode ini memberikan keuntungan karena hasil analisis yang diperlukan dapat diperoleh relatif lebih cepat dibandingkan metode konvensional lainnya.

2.6 RINGKASAN

Dalam bab ini menyajikan pendekatan D_ELT yang kami usulkan untuk mentransformasi dan menganalisis data secara efisien, sehingga membuatnya dapat digunakan untuk sejumlah besar data sensor besar, terutama data besar geografis. Berdasarkan hasil percobaan, kami melakukan beberapa observasi sebagai berikut. Pertama, D_ELT mengungguli ETL dan ELT selama persiapan data. Kedua, D_ELT menunjukkan penurunan kinerja selama analisis data. Namun, semakin tinggi kompleksitas analisisnya, semakin kecil penurunan kinerjanya, sehingga menghasilkan peningkatan kinerja secara keseluruhan dibandingkan dengan ETL atau ELT. Terakhir, dalam kasus analisis sederhana yang meningkatkan cakupan pengoptimalan, D_ELT yang dioptimalkan mengungguli ELT. Di masa depan, kami berencana untuk lebih meningkatkan kinerja keseluruhan sistem yang kami kembangkan termasuk D_ELT dan Marmot dengan menyelidiki indeks spasial, untuk lebih mendukung kueri spasial dalam menangani data besar geografis.

BAB 3
ANALISIS OVERLAY DAN POLIGON GEOGRAFIS DALAM CLOUD

Analisis overlay adalah tugas umum dalam komputasi geografis yang banyak digunakan dalam sistem informasi geografis, grafik komputer, dan ilmu komputer. Dengan adanya terobosan dalam teknologi pengamatan Bumi, khususnya munculnya teknologi penginderaan jarak jauh satelit resolusi tinggi, data geografis telah menunjukkan pertumbuhan yang luar biasa. Analisis overlay data geografis yang masif dan kompleks telah menjadi tugas komputasi yang intensif. Pemrosesan paralel terdistribusi di lingkungan cloud memberikan solusi efisien untuk masalah ini. Paradigma komputasi awan yang diwakili oleh Spark telah menjadi standar untuk pemrosesan data besar-besaran di industri dan akademisi karena karakteristiknya yang berskala besar dan latensi rendah. Paradigma komputasi awan telah menarik perhatian lebih lanjut untuk tujuan memecahkan analisis overlay data yang sangat besar. Studi-studi ini terutama berfokus pada bagaimana menerapkan analisis overlay paralel dalam paradigma komputasi awan namun kurang memperhatikan dampak kompleksitas grafik data spasial terhadap efisiensi komputasi paralel, terutama kemiringan data yang disebabkan oleh perbedaan kompleksitas grafis. Poligon geografis seringkali memiliki struktur grafis yang kompleks, seperti banyak simpul, struktur komposit termasuk lubang dan pulau. Ketika paradigma Spark digunakan untuk menyelesaikan analisis overlay poligon geografis yang masif, efisiensi penghitungannya berkaitan erat dengan faktor-faktor seperti organisasi data dan desain algoritme. Mengingat pengaruh kompleksitas bentuk poligon terhadap kinerja analisis overlay, kami merancang dan mengimplementasikan algoritma pemrosesan paralel berdasarkan paradigma Spark dalam buku ini ini. Berdasarkan analisis kompleksitas bentuk poligon, kecepatan analisis overlay ditingkatkan melalui partisi data yang masuk akal, indeks spasial terdistribusi, filter persegi batas minimum dan proses optimasi lainnya, serta kecepatan tinggi dan efisiensi paralel dipertahankan.

3.1 ANALISIS OVERLAY

Analisis overlay adalah operasi komputasi geografis yang umum dan fungsi analisis spasial yang penting dari sistem informasi geografis (GIS). Ini banyak digunakan dalam aplikasi yang berkaitan dengan komputasi spasial. Operasi ini melibatkan analisis overlay spasial dari berbagai lapisan data dan atributnya di area target. Ini menghubungkan beberapa objek spasial dari beberapa kumpulan data, membuat kumpulan data klip baru, dan menganalisis secara kuantitatif rentang spasial dan karakteristik interaksi antara berbagai jenis objek spasial. Perkembangan ilmu geografis telah memasuki babak baru dengan pesatnya mempopulerkan Internet global, teknologi sensor, dan teknologi observasi Bumi. Transformasi layanan informasi ruang angkasa dari Bumi digital ke Bumi cerdas telah menimbulkan tantangan, seperti padatnya data, intensif komputasi, padat waktu-ruang, dan akses bersamaan yang tinggi. Analisis overlay berkaitan dengan data yang sangat besar, yang mana algoritma dan model pemrosesan data tradisional tidak lagi cocok. Misalnya, jumlah

petak klasifikasi penggunaan lahan di Provinsi Yunnan yang diteliti dalam penelitian ini berjumlah ratusan ribu di tingkat kabupaten, jutaan di tingkat kota, dan puluhan juta di tingkat provinsi. Dengan berkembangnya ekonomi sosial dan kemajuan teknologi akuisisi data, jumlah patch klasifikasi penggunaan lahan akan terus meningkat. Menghitung perubahan penggunaan lahan secara efektif menggunakan model perhitungan komputer tunggal tradisional adalah hal yang sulit.

Munculnya teknologi komputasi paralel, seperti pengelompokan jaringan, komputasi grid, dan pemrosesan terdistribusi dalam beberapa tahun terakhir telah secara bertahap menggeser penelitian komputasi spasial GIS berkinerja tinggi dari optimalisasi algoritma ke transformasi paralel dan desain strategi paralel komputasi spasial GIS di lingkungan komputasi awan. Baru-baru ini, teknologi MapReduce dan Spark telah diterapkan pada analisis overlay data spasial yang sangat besar, dan beberapa hasil telah dicapai. Namun demikian, data spasial masif berbeda dengan data internet masif pada umumnya. Karakteristik spasial data spasial dan kompleksitas algoritma analisis spasial menentukan bahwa hanya menyalin paradigma pemrograman komputasi awan tidak dapat mencapai komputasi geografis berkinerja tinggi. Oleh karena itu, penelitian ini memilih algoritma kliping Hormann klasik untuk menganalisis dan mengukur dampak kompleksitas bentuk poligon geografis pada analisis overlay paralel, dan mengusulkan metode partisi Hilbert berdasarkan ukuran kompleksitas bentuk untuk mengatasi kemiringan data yang disebabkan oleh perbedaan kompleksitas bentuk poligon. Selain itu, melalui kombinasi pemfilteran MBR (Minimum Bounding Rectangle), indeks spasial R-tree, dan optimalisasi lainnya, algoritma analisis overlay paralel yang efisien dirancang. Analisis eksperimental menunjukkan bahwa metode yang diusulkan mengurangi jumlah operasi persimpangan poligon, mencapai penyeimbangan beban tugas komputasi yang lebih baik, dan sangat meningkatkan efisiensi paralel analisis overlay. Ketika inti komputasi meningkat, algoritme mencapai rasio akselerasi ke atas, dan kinerja komputasi menunjukkan perubahan nonlinier.

3.2 PEKERJAAN YANG RELEVAN

Buku ini membahas pekerjaan penelitian terkait dari dua aspek: kompleksitas bentuk dan algoritma analisis overlay.

Kompleksitas Bentuk

Banyak penelitian menggunakan bahasa abstrak untuk mendeskripsikan bentuk dan detail kompleks objek geometris, seperti “struktur poligon dengan banyak lubang, jumlah simpul sangat banyak, dan poligon dengan banyak cekung”. Untuk mengevaluasi biaya komputasi, kompleksitas dan efisiensi komputasi masalah komputasi geometri harus diukur secara akurat. Banyak aplikasi yang berkaitan dengan komputasi spasial sangat bergantung pada algoritma untuk memecahkan masalah geometri. Ketika berhadapan dengan masalah komputasi geografis skala besar, evaluasi biaya komputasi harus mempertimbangkan kuantitas data input, kompleksitas objek grafis, dan kompleksitas waktu model komputasi. Ketika jumlah data masukan dan algoritme ditentukan, kompleksitas objek grafis yang berbeda sering kali menyebabkan perbedaan besar dalam efisiensi komputasi.

Mandelbrot menggambarkan kompleksitas objek geometris dari perspektif dimensi fraktal. Metode yang paling umum digunakan adalah teknik penghitungan kotak. Brinkhoff secara kuantitatif melaporkan kompleksitas poligon dari tiga aspek, yaitu frekuensi getaran lokal, amplitudo getaran lokal, dan deviasi dari convexhull, untuk menggambarkan kompleksitas bentuk global. Berdasarkan penelitian Brinkhoff, Bryson mengusulkan kerangka konseptual untuk membahas ukuran kompleksitas bentuk berorientasi pemrosesan kueri untuk objek spasial. Rossignac menganalisis kompleksitas bentuk dari aspek kompleksitas aljabar, topologi, morfologi, kombinatorial, dan ekspresi. Rossignac juga mengurangi kompleksitas bentuk dengan menggunakan representasi batas segitiga pada skala berbeda. Ying mengoptimalkan transmisi data grafis berdasarkan kompleksitas bentuk.

Dari pembahasan di atas, kita mengetahui bahwa kompleksitas grafis mempunyai arti dan metode pengukuran yang berbeda-beda di berbagai bidang profesional, seperti kompleksitas desain, kompleksitas visual, dan sebagainya. Oleh karena itu, kita harus mempertimbangkan kompleksitas bentuk dari perspektif komputasi geografis. Kompleksitas bentuk secara langsung mempengaruhi efisiensi analisis spasial dan perhitungan kueri spasial, seperti jumlah simpul dan bentuk lokal (seperti cekungan) grafik, sehingga sangat mempengaruhi efisiensi perhitungan geometri spasial. Nilai-nilai ini merupakan indikator penting untuk mengevaluasi perhitungan biaya. Mempertimbangkan sepenuhnya pengaruh kompleksitas grafis pada komputasi geografis tertentu dapat mengoptimalkan efisiensi komputasi aplikasi secara efektif.

Analisis Overlay

Kajian tentang analisis aritmatika vektor overlay bermula dari bidang grafik komputer. Misalnya, dua kelompok yang terdiri dari ribuan poligon hamparan sering kali terpotong dalam rendering grafis 2D dan 3D. Selanjutnya, algoritma analisis overlay yang berbeda telah dihasilkan. Di antaranya, algoritma Sutherland–Hodgman, Vatti, dan Greiner–Hormann adalah yang paling representatif ketika menangani kliping poligon sembarang. Algoritma Sutherland – Hodgman tidak cocok untuk poligon kompleks. Algoritme Weiler – Atherton mengharuskan calon poligon disusun searah jarum jam dan tanpa poligon yang berpotongan sendiri. Algoritme Vatti tidak membatasi jenis kliping, sehingga poligon yang berpotongan sendiri dan berpori juga dapat diproses. Algoritma Hormann memotong poligon dengan menilai pintu masuk dan keluar garis arah. Algoritma ini juga mengatasi degradasi titik dengan memindahkan jarak kecil. Selain itu, algoritma Hormann dapat menangani poligon yang berpotongan sendiri dan noncembung. Algoritma Weiler menggunakan struktur data pohon, sedangkan algoritma Vatti dan Greiner – Hormann mengadopsi struktur data daftar tertaut bilinear. Oleh karena itu, algoritma Vatti dan Greiner–Hormann lebih baik dibandingkan algoritma Weiler dalam hal kompleksitas dan kecepatan lari.

Peneliti selanjutnya telah menerapkan banyak perbaikan pada algoritma kliping vektor tradisional yang disebutkan di atas, yang menyederhanakan penghitungan kliping poligon vektor. Namun, penelitian ini didasarkan pada optimasi algoritma serial. Ketika analisis overlay diterapkan pada bidang komputasi geografis, analisis tersebut akan menangani poligon yang lebih kompleks (seperti poligon berlubang dan pulau) dan volume data yang lebih besar

(jumlah patch klasifikasi penggunaan lahan di suatu provinsi adalah puluhan atau bahkan ratusan patch). jutaan, dan sebuah poligon mungkin memiliki puluhan ribu simpul). Algoritme kliping vektor dapat diterapkan secara efisien pada grafik komputer tetapi tidak dapat diterapkan secara efisien pada komputasi geografis. Selain itu, banyak algoritma kliping elemen geografis tradisional juga menunjukkan kesesuaian yang buruk dan penurunan kinerja. Dengan perkembangan teknologi komputer dan peningkatan volume data spasial, algoritma kliping vektor tradisional sering kali menghadapi hambatan efisiensi ketika berhadapan dengan kumpulan data geografis yang besar dan kompleks. Oleh karena itu, meningkatkan algoritma overlay dan menggunakan platform komputasi paralel untuk analisis overlay data masif merupakan arah penelitian baru.

Dengan pesatnya perkembangan teknologi komputasi awan MapReduce dan Spark, penggunaan penyimpanan terdistribusi berskala besar dan teknologi komputasi paralel untuk pemrosesan dan analisis data dalam jumlah besar telah menjadi pendekatan teknis yang efektif. Studi terbaru telah menerapkan teknologi MapReduce dan Spark pada analisis overlay data spasial yang sangat besar. Wang menggunakan MapReduce untuk meningkatkan efisiensi analisis overlay sekitar 10 kali lipat dengan partisi grid dan indeks. Zheng membangun struktur indeks grid bertingkat dengan menggabungkan grid tingkat pertama dengan quartering berdasarkan platform komputasi terdistribusi Spark. Eksperimen Zheng menunjukkan bahwa algoritma indeks grid mencapai hasil yang baik ketika poligon terdistribusi secara merata; jika tidak, efisiensi algoritmanya rendah. Xiao membuktikan bahwa partisi tugas paralel berdasarkan lokasi spasial poligon menghasilkan penyeimbangan beban yang lebih baik daripada partisi tugas acak. Selain itu, SpatialHadoop dan GeoSpark memperluas Hadoop dan Spark untuk mendukung komputasi data spasial masif dengan lebih baik. Diantaranya, Spatial Hadoop merancang seperangkat model penyimpanan objek spasial, yang menyediakan HDFS dengan grid, R-tree, kurva Hilbert, kurva Z dan indeks lainnya. Selain itu, juga menyediakan fungsi filtering untuk memfilter data yang tidak perlu diproses. GeoSpark juga menambahkan serangkaian model objek spasial dan memperluas RDD (Resilient Distributed Dataset) ke SRDD (Spatial Resilient Distributed Dataset) yang mendukung penyimpanan objek spasial. GeoSpark juga menyediakan fungsi pemfilteran untuk menyaring data yang tidak perlu diproses. Ide desain SpatialHadoop dan GeoSpark memiliki nilai referensi yang besar untuk buku ini.

Singkatnya, menggunakan paradigma komputasi awan Spark untuk mengembangkan komputasi geografis berkinerja tinggi adalah metode yang murah dan berkinerja tinggi. Ini juga merupakan salah satu pusat penelitian di bidang komputasi geografis berkinerja tinggi. Studi terbaru telah menerapkan analisis overlay di Spark, yang sangat meningkatkan efisiensi analisis overlay. Mengoptimalkan strategi untuk karakteristik data spasial, seperti partisi data yang masuk akal dan indeks data spasial yang sangat baik, memainkan peran penting dalam meningkatkan efisiensi komputasi paralel, dan partisi Hilbert lebih cocok untuk analisis overlay paralel pada data distribusi spasial yang tidak seragam dibandingkan grid partisi. Terlihat bahwa analisis overlay paralel saat ini didasarkan pada antarmuka kliping pihak ketiga dan mengabaikan dampak kompleksitas bentuk poligon geografis pada algoritma kliping, yang akan menyebabkan ketidakseimbangan data yang serius.

3.3 ALGORITMA ANALISIS OVERLAY

Pada bagian ini, ide inti pada bab ini akan diperkenalkan. Pertama, algoritma analisis overlay dasar yang sangat baik dipilih untuk dieksekusi pada setiap node komputasi. Kemudian, menurut kompleksitas dan lokasi grafik, poligon dibagi secara wajar, dan indeks berdasarkan lokasi spasial dibuat untuk mewujudkan akses data yang cepat dan penyeimbangan beban node komputasi paralel.

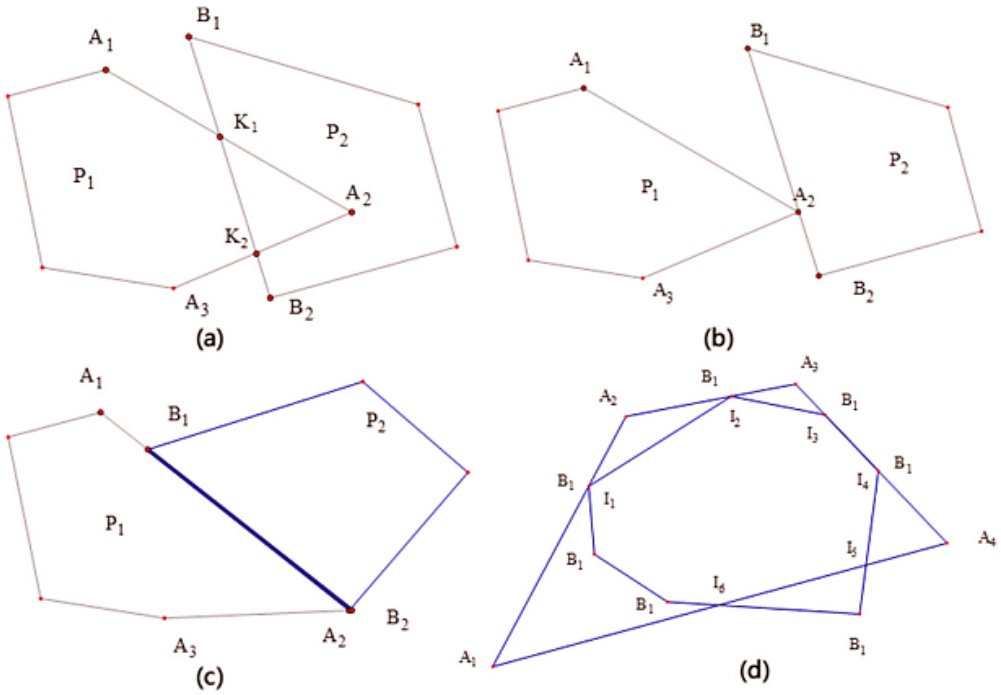
Algoritma Analisis Overlay Dasar yang Berjalan pada Setiap Node Komputasi.

Algoritma analisis overlay dasar, yang merupakan program pemrosesan dasar untuk setiap node komputasi paralel, melakukan analisis overlay pada dua set poligon. Dalam merancang algoritma analisis overlay, kami menggunakan algoritma Hormann, yang dapat menangani struktur kompleks (misalnya, perpotongan diri dan polimorfisme berlubang), sebagai referensi. Namun, metode perturbasi koordinat titik yang digunakan oleh algoritma Hormann bukanlah solusi terbaik untuk masalah degradasi, yang menyebabkan kesalahan kumulatif dalam statistik area patch masif. Oleh karena itu, kami menyelesaikan degenerasi perpotongan dengan menilai interval azimuth antara garis korelasi yang berpotongan. Kami juga menggunakan algoritma yang ditingkatkan sebagai dasar analisis overlay paralel.

Untuk mencapai ekspresi yang disederhanakan dari analisis overlay dua kelompok poligon, kita asumsikan bahwa setiap kelompok poligon hanya memiliki satu objek poligon, karena analisis overlay dua kelompok lapisan dengan banyak poligon hanya meningkatkan jumlah iterasi. Langkah-langkah pemrosesan algoritma Hormann yang ditingkatkan adalah sebagai berikut:

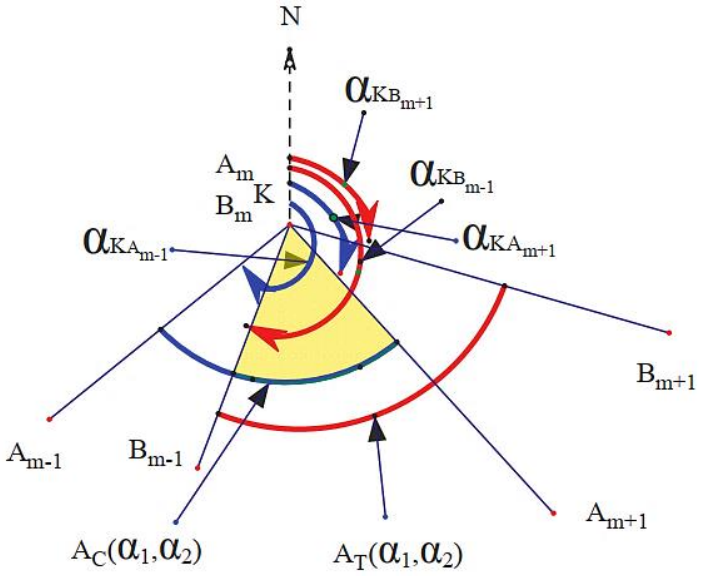
1. Menghitung perpotongan poligon yang terpotong dan poligon target
2. Menilai masuk dan keluar titik potong dengan ruas garis vektor (menilai titik masuk atau keluar titik potong) dan menjumlahkan titik masuk pada barisan titik sudut poligon hasil kliping
3. Membandingkan interval azimuth dari simpul-simpul yang mengalami degenerasi dari titik-titik perpotongan dan menambahkan simpul-simpul yang tumpang tindih dari interval azimuth ke barisan simpul dari poligon hasil kliping
4. Membentuk poligon baru (hasil kliping) sesuai dengan barisan simpulnya

Seperti yang ditunjukkan pada Gambar 4.1a, poligon P_1 yang terpotong dan poligon target P_2 berpotongan. Titik potong K_1 dan K_2 dapat diperoleh melalui persamaan kolinearitas. Dengan menilai nilai positif dan negatif hasil kali ruas garis vektor, titik potong masuk dan keluar dapat dinilai. Seperti diilustrasikan pada gambar yang sama, $\overrightarrow{A_1A_2} \times \overrightarrow{B_1B_2} > 0$, dan dengan demikian, K_1 adalah titik masuk relatif terhadap P_2. Selain itu, $\overrightarrow{A_2A_3} \times \overrightarrow{B_1B_2} < 0$,, sehingga K_2 adalah titik keluar. Poligon yang dihasilkan terdiri dari barisan simpul yang terdiri dari K_1, A_2, dan K_2. Seperti diilustrasikan pada Gambar 3.1b–d, titik masuk dan titik keluar tidak sesuai untuk menggambarkan degradasi persimpangan.



Gambar 3.1. Hamparan poligon.

Gambar 3.2 menunjukkan cara menangani fenomena degenerasi persimpangan. Pada Gambar 3.2, panah putus-putus N menunjuk ke arah utara, yang merupakan titik awal perhitungan azimuth. Oleh karena itu, setiap ruas garis memiliki azimuthnya masing-masing. Poligon terpotong dan poligon sasaran mempunyai titik potong K yang juga merupakan letak simpul A_m dan B_m. Interval azimuth dari poligon terpotong dan poligon target di persimpangan K adalah $\mathrm{A_C}(\alpha_1,\alpha_2)$ dan $\mathrm{A_T}(\alpha_1,\alpha_2)$. Jika $\mathrm{A_C}(\alpha_1,\alpha_2)$ dan $\mathrm{A_T}(\alpha_1,\alpha_2)$. mempunyai bagian yang bertumpang tindih (warna kuning pada gambar), maka kedua poligon bertumpang tindih di dekat titik potong, sehingga harus ditambahkan ke barisan titik sudut dari poligon yang dihasilkan.



Gambar 3.2. Diagram perhitungan interval azimuth.

Pengaruh Kompleksitas Bentuk pada Efisiensi Kliping Paralel

Dalam komputasi kliping paralel, setiap node komputasi biasanya diberi jumlah poligon yang sama. Secara umum, sulit untuk memastikan bahwa poligon kompleks dialokasikan secara merata ke setiap node komputasi; biasanya, satu node komputasi dialokasikan poligon yang lebih kompleks. Meskipun jumlah total poligon yang dialokasikan oleh setiap node komputasi sama, namun node komputasi ini memerlukan waktu yang lama untuk menyelesaikan tugas komputasi yang dialokasikan, sedangkan node komputasi lainnya akan berada dalam keadaan menunggu. Oleh karena itu, mengabaikan perbedaan kompleksitas poligon akan mengakibatkan situasi di mana setiap node komputasi tidak dapat menyelesaikan tugas komputasi pada saat yang sama, sehingga efisiensi komputasi paralel berkurang.

Kompleksitas adalah konsep linguistik intuitif. Secara umum, bidang profesional yang berbeda memberikan perhatian yang berbeda terhadap kompleksitas bentuk. Dalam bidang komputasi geografis, kompleksitas bentuk berhubungan dengan algoritma geografis tertentu. Bentuk yang sama berhubungan dengan algoritma geografis yang berbeda dan mungkin mempunyai kompleksitas bentuk yang berbeda.

Di sisi lain, khusus poligon geografis, poligon yang berbeda mempunyai ciri morfologi yang berbeda pula, seperti cembung, cekung, perpotongan sendiri, dan jumlah simpul yang banyak. Untuk mengukur kompleksitas, informasi harus dikompresi menjadi satu atau lebih parameter dan model ekspresi yang sebanding. Meskipun titik awal dan lokasinya sangat berbeda, bentuk serupa mungkin masih muncul. Oleh karena itu, ketika membahas kompleksitas bentuk poligon, kita dapat mengabaikan posisi dan skala spasial, dan fokus pada pengaruh fitur poligon pada komputasi geografis.

Kompleksitas bentuk dapat didefinisikan dari perspektif komputasi geografis:

Definisi 1. Kompleksitas bentuk adalah ukuran indeks intensitas komputasi dari bentuk yang berpartisipasi dalam perhitungan algoritma geografis. Kompleksitas bentuk dapat diukur dengan banyaknya pengulangan operasi dasar dalam suatu algoritma geografis yang disebabkan oleh suatu bentuk. Sedangkan untuk analisis overlay poligon, operasi paling dasar dari algoritma Hormann adalah mencari titik potong dua sisi. Oleh karena itu, untuk algoritma Hormann, kompleksitas suatu poligon adalah jumlah sisi yang dimilikinya. Berdasarkan analisis tersebut, kita mengetahui bahwa kompleksitas bentuk merupakan nilai absolut yang sulit diprogram. Oleh karena itu, perlu diperoleh nilai relatif melalui normalisasi untuk mengukur kompleksitas grafis.

Definisi 2. Diketahui himpunan poligon $P = (P_1, P_2, \dots P_n)$banyaknya simpul pada poligon tersebut adalah V_i, V_{min} adalah banyaknya simpul minimum seluruh poligon, dan V_{max} adalah banyaknya simpul maksimum simpul dari semua poligon. Kemudian kompleksitas Wi dari poligon P_i dapat dinyatakan sebagai (Persamaan 1)

$$W_i = \frac{V_i - V_{min}}{V_{max} - V_{min}}$$

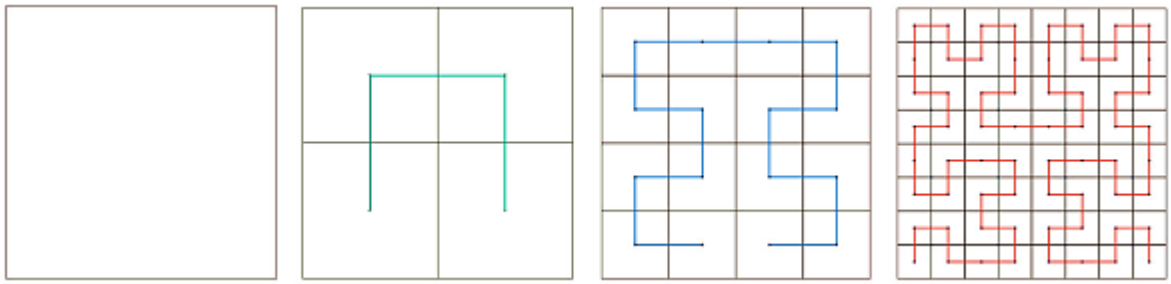
Karena poligon biasanya diwakili oleh barisan simpul dalam model penyimpanan poligon. Jumlah sisi suatu poligon sama dengan jumlah simpul, jadi pada Definisi 2, kita menggunakan simpul suatu poligon, bukan sisi. Oleh karena itu, dalam analisis overlay paralel, kita dapat menjadikan kompleksitas bentuk sebagai indikator partisi data.

Keadaan idealnya adalah kompleksitas poligon setiap partisi data adalah sama, di mana semua node komputasi akan menyelesaikan tugas komputasi pada waktu yang sama.

3.4 METODE PENYEIMBANGAN DAN PARTISI DATA

Partisi data adalah kunci untuk mempercepat algoritma kliping poligon berdasarkan platform komputasi berkinerja tinggi. Sepotong data lengkap dibagi menjadi data multiblok independen yang relatif kecil, yang menyediakan dasar untuk operasi data terdistribusi atau paralel. Partisi data spasial berbeda dengan partisi data umum. Selain menyeimbangkan jumlah data, hubungan lokasi spasial, seperti agregasi spasial dan kedekatan data, juga harus dipertimbangkan. Metode partisi data spasial yang umum digunakan adalah partisi meshing dan filling curve. Meshing sederhana dan mempertimbangkan kedekatan spasial data, namun tidak dapat menjamin jumlah data yang seimbang. Kurva Hilbert merupakan kurva pengisian spasial klasik dengan karakteristik pengelompokan spasial yang baik dan mempertimbangkan hubungan spasial dan muatan data. Oleh karena itu, strategi partisi data pada penelitian ini mengadopsi algoritma kurva pengisian Hilbert yang dipadukan dengan kompleksitas bentuk untuk mencapai penyeimbangan beban.

Pada Gambar 2.3, partisi Hilbert membagi wilayah spasial menjadi grid 2N × 2N. Selama proses iterasi, N adalah orde kurva Hilbert, yaitu jumlah iterasi. Secara umum N ditentukan oleh jumlah objek spasial, dan jumlah data spasial membutuhkan $n < 22 \times N$.



Gambar 3.3. Partisi Hilbert dan pembuatan kurva Hilbert.

Partisi Hilbert terdiri dari empat langkah berikut:

1. Tentukan orde kurva Hilbert, buat grid Hilbert dan kurva Hilbert, beri nomor pada kurva Hilbert secara berurutan, dan dapatkan himpunan pengkodean grid Hilbert, $GHid = \{GH_1, GH_2 \cdots GH_n\}$.
2. Hitung titik pusat MBR poligon, temukan mesh yang sesuai, dan gunakan pengkodean Hilbert dari mesh tersebut sebagai pengkodean Hilbert dari poligon untuk mendapatkan himpunan pengkodean Hilbert dari poligon, $PHid = \{PH_1, PH_2 \cdots PH_n\}$.

3. Sesuai dengan jumlah node komputasi M, bagilah himpunan pengkodean Hilbert dari poligon menjadi M partisi, dan hitung pengkodean start-stop dari pengkodean poligon Hilbert di setiap partisi.
   Gabungkan grid partisi Hilbert untuk mendapatkan poligon partisi $PS = \{PS_1, PS_2, \cdots, PS_M\}$.
4. Dalam partisi sebenarnya, bentuk poligon berbeda karena poligon tidak berada dalam distribusi seragam yang ideal. Jika hanya satu titik pusat poligon yang diperlukan untuk setiap grid, maka orde Hilbert N mungkin sangat besar, panjang tepi grid akan terlalu kecil, dan tidak ada pusat MBR poligonal yang mungkin ada di banyak grid. Jadi, partisi Hilbert dan kurva Hilbert akan memakan banyak waktu komputasi, dan perhitungan overlay selanjutnya akan melibatkan banyak masalah lintas partisi. Oleh karena itu, keberadaan beberapa pusat MBR poligonal dalam sebuah grid diperlukan.

Urutan N dari grid Hilbert berhubungan dengan panjang tepi mesh. Panjang grid dan orde N juga ditentukan. Untuk mendapatkan urutan N yang masuk akal pada kurva Hilbert, kita dapat menghitung distribusi normal posisi titik pusat poligon MBR, menentukan panjang tepi kisi yang optimal, dan pada akhirnya mencapai keseimbangan antara urutan N pada kurva Hilbert dan jumlah poligon titik pusat MBR di setiap grid. Kunci untuk membagi *PHid* dari kumpulan poligon pengkodean Hilbert adalah memastikan keseimbangan beban setiap partisi. Mengingat kompleksitas poligon dapat sangat bervariasi, kita tidak bisa begitu saja membagi kumpulan kode Hilbert PH_i dari poligon secara merata.

Kompleksitas bentuk poligon Pi didefinisikan sebagai Wi., W sebagai kompleksitas rata-rata semua poligon, kompleksitas ideal setiap partisi sebagai Wideal, dan kompleksitas aktual sebagai Wactual, maka (Persamaan 2)

$$W_{ideal} = \frac{\sum_{i=1}^{n} W_i}{M}$$

jika poligon dari j sampai k ditempatkan pada partisi yang sama, maka,
(Pertemuan 3)

$$W_{actual} = \sum_{i=j}^{k} W_i$$

(Persamaan 4)

$$|W_{ideal} - W_{actual}| < \bar{W}$$

Secara umum, jumlah poligon di setiap partisi sedikit berbeda setelah dipartisi, namun kompleksitas poligon di setiap partisi pada dasarnya sama. Oleh karena itu, strategi ini menjamin penyeimbangan beban tugas komputasi.

Indeks R-tree

R-tree adalah metode indeks data spasial yang diadopsi secara luas; ini digunakan dalam perangkat lunak komersial, seperti Oracle dan SQL Server. Untuk meningkatkan efisiensi akses data spasial, harus dibangun R-tree. Selain itu, data disegmentasi sesuai dengan titik partisi data Hilbert, dan area grid dari kurva Hilbert sebelum setiap titik partisi

didefinisikan sebagai area sub-indeks. Selain itu, indeks R-tree untuk objek spasial ditetapkan di wilayah sub-indeks. Demikian pula, hubungan pemetaan antara pengkodean grid, pengkodean titik pusat MBR poligon, dan pengkodean area sub-indeks juga dibuat. Selanjutnya, kode indeks yang sesuai pada node komputasi di-cache. Dalam percobaan ini, kami secara langsung menggunakan kelas STR-tree (Sort Tile Recursive) dari perpustakaan JTS (Java Topology Suite, perpustakaan perangkat lunak java) untuk membuat indeks R-tree.

3.5. DESAIN PROSES ANALISIS OVERLAY PARALEL TERDISTRIBUSI

Untuk memastikan bahwa proses ini cocok untuk decoupling, kami membagi proses analisis overlay paralel terdistribusi menjadi enam langkah berdasarkan karakteristik algoritme: Pemrosesan awal data, pemfilteran awal, pengkodean Hilbert, partisi data dan pembuatan indeks, pemfilteran data, dan perhitungan overlay (Gambar 3.4).

(1) Pemrosesan awal data

Dalam keseluruhan proses, banyak langkah yang perlu dilakukan untuk melintasi semua poligon dan simpulnya. Untuk mengurangi jumlah traversal, kita dapat melakukan pemrosesan terpusat dalam satu traversal, seperti menghitung MBR suatu poligon, titik pusat geometrinya, luas MBR, dan kompleksitas bentuk poligon, untuk menyiapkan data guna mengoptimalkan traversal. aliran pemrosesan. Pada perhitungan selanjutnya, informasi tersebut dapat langsung dibaca untuk menghindari perhitungan berulang. Mengingat banyaknya data dalam prapemrosesan, paradigma Spark dapat digunakan dalam pemrosesan data paralel. Sesuai dengan perhitungan jumlah node fisik N, data dibagi secara default, dan data yang dialokasikan dilintasi pada setiap node komputasi.

(2) Penyaringan awal

Dapat ditentukan bahwa hanya poligon di area perpotongan MBR dari dua lapisan poligonal yang perlu dipotong. Oleh karena itu, memfilter poligon yang tidak memerlukan kliping dapat mengurangi biaya komputasi.

(3) Pengkodean Hilbert dan penghitungan titik partisi data berdasarkan kompleksitas klip poligon

Semua poligon dibagi oleh grid Hilbert sesuai dengan posisi distribusi spasial, dan setiap grid dan poligon diberi kode Hilbert. Kemudian, titik partisi data dihitung berdasarkan kompleksitas bentuk. Pekerjaan ini tidak cocok untuk decoupling dan, oleh karena itu, tidak dapat dilaksanakan secara paralel.

(4) Partisi dan pengindeksan data

Berdasarkan titik partisinya, wilayah kurva Hilbert dianggap sebagai wilayah sub-indeks. Kelas STR-tree dari perpustakaan JTS digunakan untuk membuat indeks R-tree untuk setiap partisi, dan file indeks disimpan di setiap node komputasi.

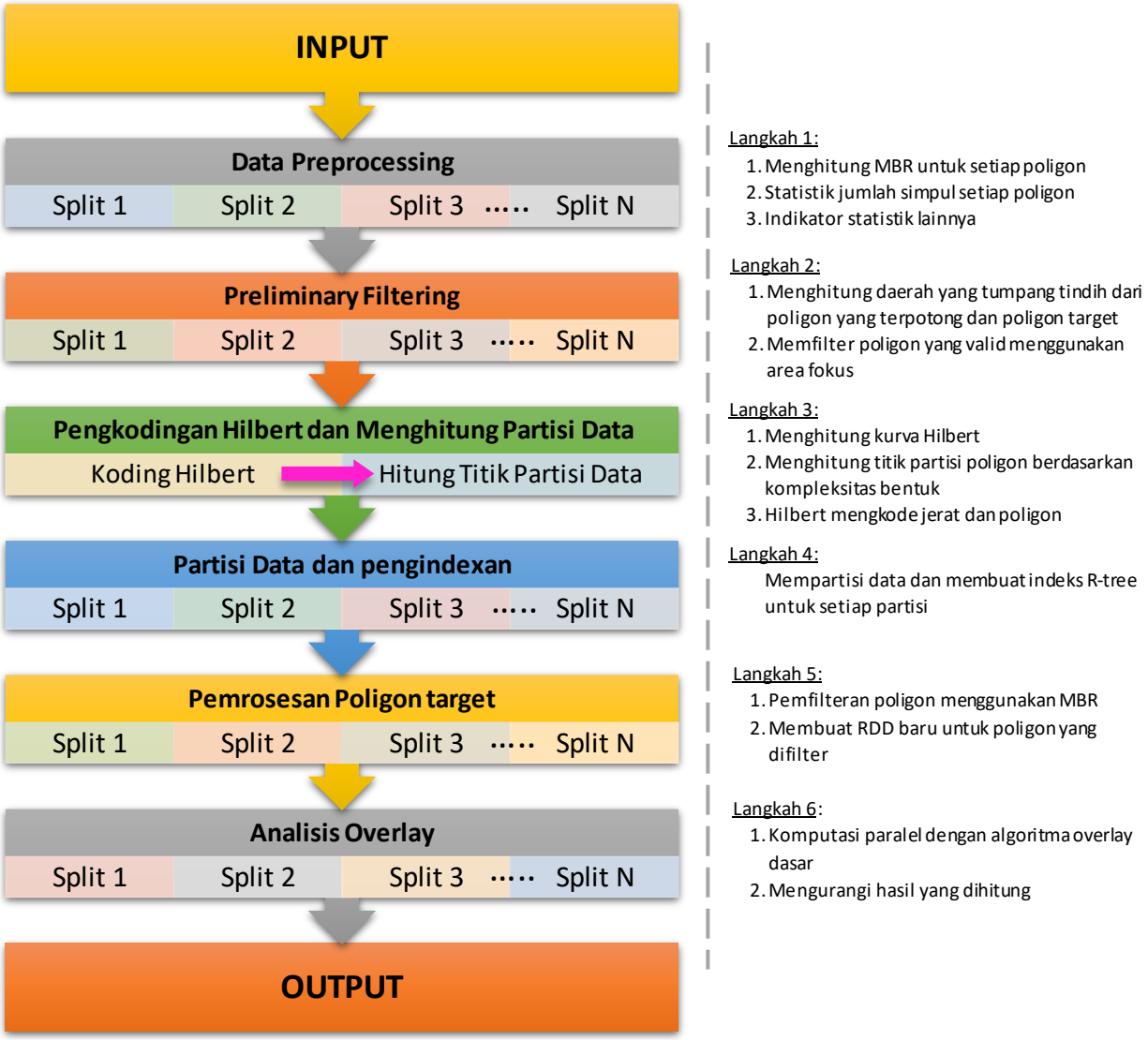
(5) Pemfilteran poligon target

Dalam analisis overlay, setiap titik dari poligon yang terpotong dan poligon target harus dilintasi. Bahkan jika kedua poligon tidak tercakup, semua titik akan dilintasi, sehingga menghasilkan beberapa perhitungan yang tidak valid. Menyaring poligon

yang tidak valid dari poligon target jelas dapat meningkatkan efisiensi. Sebelum perhitungan overlay, poligon target tanpa analisis overlay dapat dihilangkan secara efektif dengan menghitung apakah ada hubungan overlay antara MBR dari poligon yang terpotong dan MBR dari poligon target. Metode perhitungannya secara langsung membandingkan koordinat maksimum dan minimum dari poligon yang terpotong dan poligon target tanpa menggunakan algoritma overlay.

(6) Analisis hamparan

Semua node komputasi menggunakan algoritma Hormann untuk komputasi overlay paralel. Hasil setiap node perhitungan direduksi untuk mendapatkan hasil analisis overlay akhir.



Gambar 3.4. Aliran komputasi overlay paralel.

Analisis Algoritma

Proses utama analisis overlay paralel yang dilakukan dalam penelitian ini meliputi pemrosesan awal data, pemfilteran awal, partisi Hilbert, penetapan indeks R-tree, pemfilteran MBR poligon, dan kliping poligon.

Dalam prapemrosesan data, hanya data atribut lapisan dan informasi koordinat titik poligon yang disertakan dalam data asli. Poligon MBR dan jumlah simpul poligon harus digunakan tiga kali dalam proses perhitungan yang dirancang dalam penelitian ini. Oleh karena itu, kami menyatukan prapemrosesan data, membuat struktur data baru, dan menghindari penghitungan berulang. Pemrosesan awal data terpadu menghemat sekitar setengah beban kerja dibandingkan dengan pemrosesan awal data terpisah.

Pada Preliminary filtering, kompleksitas waktu algoritma filtering MBR adalah O(1), sedangkan kompleksitas algoritma analisis overlay adalah O(logN), dimana N adalah jumlah simpul poligon. Oleh karena itu, kompleksitas komputasi akan sangat berkurang dengan memfilter poligon tanpa analisis overlay melalui MBR. Selain itu, berkurangnya kompleksitas komputasi bergantung pada distribusi spasial poligon, yang merupakan faktor yang tidak dapat dikendalikan.

Kompleksitas waktu dalam membangun kurva Hilbert adalah $O(N)^2$, dimana N adalah orde kurva Hilbert. Semakin besar N maka semakin lama waktu yang dibutuhkan untuk mempartisi data. Namun, jika N terlalu kecil, maka beberapa poligon akan sesuai dengan pengkodean Hilbert yang sama. Jika partisi Hilbert benar-benar memenuhi kondisi bahwa setiap mesh hanya mempunyai satu titik pusat dari poligon MBR, maka grid Hilbert mendukung maksimal 22×N poligon. Selain itu, poligon dalam data nyata umumnya tidak terdistribusi secara merata, dan tidak ada poligon di banyak kisi Hilbert. Oleh karena itu, memungkinkan sejumlah nilai berulang yang dikodekan Hilbert adalah mungkin. Selain itu, dibandingkan dengan partisi grid, partisi Hilbert dapat menyelesaikan masalah ketidakrataan lokasi data dengan sempurna. R-tree adalah metode indeks data spasial yang khas. Waktu untuk traversal data dipersingkat dengan menetapkan indeks R-tree. Kompleksitas waktu R-tree adalah O(logN). Pemrosesan awal data, pemfilteran MBR, konstruksi indeks R-Tree, dan proses lainnya bersifat relatif memakan waktu. Dengan menggunakan komputasi paralel multi-node dalam partisi data, waktu yang dikonsumsi dapat dikurangi menjadi 1/N, dimana N adalah jumlah proses paralel. Dalam paradigma Spark, operasi data dilakukan di memori, dan operasi I/O memakan waktu minimal. Oleh karena itu, algoritma overlay yang diusulkan menunjukkan efisiensi yang tinggi.

3.6 DESAIN EKSPERIMENTAL

Untuk melakukan eksperimen analisis overlay, kami menggunakan petak jenis penggunaan lahan dan petak dengan kemiringan lebih dari 25 derajat di salah satu wilayah Provinsi Yunnan pada tahun 2018. Terdapat 500.000 petak jenis penggunaan lahan dan 110.000 petak lereng. Data ini tersebar di area seluas 15.000 kilometer persegi. Berdasarkan data ini, kami membuat kumpulan data berbeda untuk eksperimen. Kami akan menggunakan mode analisis overlay yang berbeda untuk eksekusi jika data besaran data berbeda, mencatat

perubahan waktu eksekusi, dan menganalisis karakteristik dan penerapan mode analisis overlay yang berbeda.

Peralatan Komputasi

Percobaan dilakukan dengan menggunakan satu komputer portabel dan enam server X86. Informasi konfigurasi peralatan ditunjukkan pada Tabel 3.1.

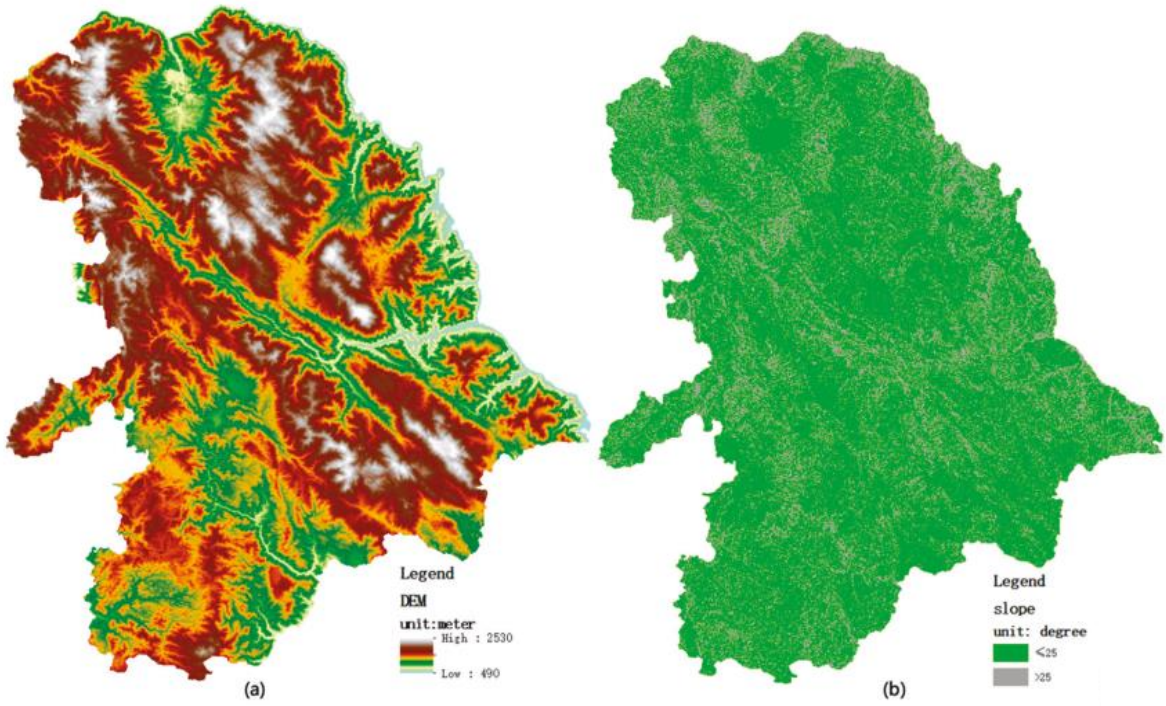
Tabel 3.1. Konfigurasi peralatan.

Peralatan	No	Konfigurasi Hardware	Operating Sistem	Software	Remark
Komputer Portabel	1	Thinkpad T470p, 8 vcore, RAM 16 G, SSD (Solid State Drive)	Windows 10	ArcMap 10.4.1	Eksperimen komputer tunggal untuk analisis overlay desktop.
X86 Server	6	DELL R720, 24 core, RAM64 G, HDD (Hard Disk Drive)	Centos7	Hadoop 2.7, Spark 2.3.1	Spark Computing Cluster

Data eksperimental

(1) Lapisan kliping

Data model elevasi digital (DEM) dari jaringan sepanjang 30 m di wilayah tersebut diperoleh dari Internet, dan peta kemiringan dihasilkan dari Internet (Gambar 3.5). Area dengan kemiringan lebih dari 25 derajat telah diekstraksi, dan diperoleh 108.025 petak.

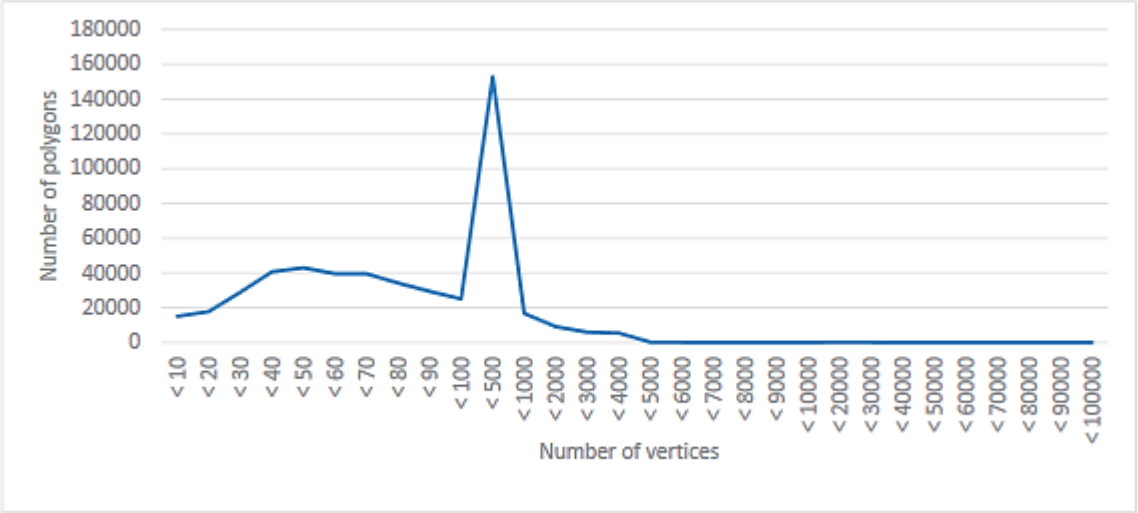


Gambar 3.5. Mengekstraksi kemiringan sebagai clipping layer dari model elevasi digital (DEM).

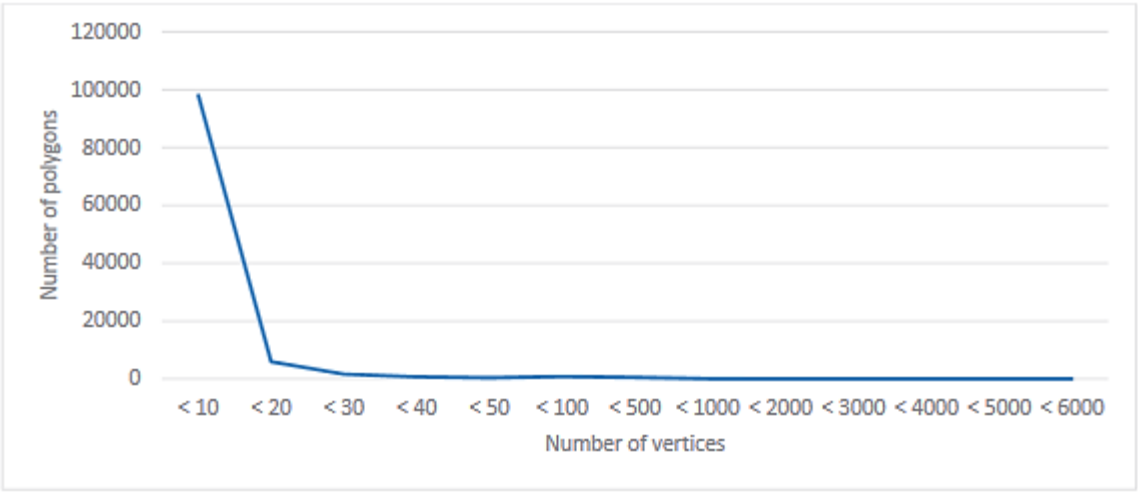
(2) Lapisan sasaran

Sebanyak 10 kelompok data eksperimen diperoleh dari 500.000 petak tipe lahan asli di wilayah tersebut dengan menggunakan kumpulan data yang jarang dan intensif. Jumlah tambalan masing-masing adalah 50.000, 100.000, 250.000, 500.000, 1 juta, 2 juta, 4 juta, 6 juta, 8 juta, dan 10 juta.

Melalui pengecekan data, ditemukan 88.000.000 simpul dalam 500.000 data pola medan asli. Di antara semua poligon, poligon paling sederhana mempunyai empat simpul, sedangkan poligon paling kompleks mempunyai 99.500 simpul. Sebanyak 890.000 simpul tercatat di 110.000 petak lereng. Di antara seluruh bidang lereng, bidang yang paling sederhana mempunyai 8 simpul, sedangkan bidang yang paling kompleks mempunyai 5.572 simpul. Statistik jumlah simpul poligon di lapisan kueri dan lapisan target diilustrasikan masing-masing pada Gambar 3.6 dan 3.7.



Gambar 3.6. Sebaran poligon dengan jumlah simpul berbeda.



Gambar 3.7. Sebaran poligon dengan jumlah simpul berbeda.

Dalam komputasi paralel, data disusun ke dalam format GeoJson dan diunggah ke HDFS. Blok data HDFS ada tiga salinan, masing-masing berukuran 64 MB.

Adegan Eksperimental

Sebelum menjelaskan desain skenario eksperimental, pertama-tama kami mendefinisikan beberapa mode berbeda untuk perbandingan dan menjelaskan perbedaan masing-masing mode (Tabel 3.2).

Tabel 3.2. Penjelasan mode eksperimen.

Mode Abbreviation	Peralatan	Mode Penyimpanan data	Catatan
ArcMap	Portabel computer dengan ArcMap	Local File System	Gunakan alat klip Toolbox untuk melakukan analisis overlay pada komputer portabel
Spark_original	Multiple X86 servers with Spark	HDFS	Langsung mempartisi data secara acak dan melakukan analisis overlay paralel tanpa perbaikan apa pun.
Spark_improved	Multiple X86 servers with Spark	HDFS	Menerapkan sepenuhnya analisis overlay paralel sesuai dengan proses Bagian 3.3. Metode partisi Hilbert mempertimbangkan kompleksitas grafik
Spark_NoComlexity	Multiple X86 servers with Spark	HDFS	Kecuali kompleksitas grafik poligon tidak dipertimbangkan, semuanya sama dengan mode Spark_improved.
Spark_MBR	Multiple X86 servers with Spark	HDFS	Berdasarkan model Spark_original, pemfilteran MBR dilakukan terlebih dahulu, kemudian dilakukan analisis overlay paralel
Spark_MBR_Hilbert	Multiple X86 servers with Spark	HDFS	Berdasarkan model Spark_original, pemfilteran MBR dan operasi partisi Hilbert ditambahkan
Spark_MBR_Hilbert_R-tree	Multiple X86 servers with Spark	HDFS	Berdasarkan model Spark_original, pemfilteran MBR, partisi Hilbert, dan operasi pembuatan indeks R-tree ditambahkan

Diantaranya, mode ArcMap adalah metode umum yang digunakan dalam pemrosesan data geografis. Tujuan membandingkan mode Spark_original, Spark_improved, dan Spark_NoComlexity adalah untuk menentukan seberapa besar peningkatan performa. Tujuan membandingkan mode pohon Spark_MBR, Spark_MBR_Hilbert dan Spark_MBR_Hilbert_R adalah untuk menentukan seberapa besar ketiga metode yang ditingkatkan dapat meningkatkan efisiensi analisis overlay paralel.

(1) Adegan 1: Bandingkan perbedaan kinerja empat mode: ArcMap, Spark_original, Spark_improved, dan Spark_NoComlexity.

Sepuluh kelompok poligon dengan nomor berbeda digunakan untuk analisis overlay dalam empat mode. Kami akan mencatat waktu penyelesaian proses analisis overlay dan menggambar kurva konsumsi waktu. Skenario eksperimental ini dapat menjawab pertanyaan-pertanyaan berikut:

- Seberapa baikkah komputasi paralel Spark dalam meningkatkan kinerja analisis overlay dibandingkan dengan perangkat lunak desktop?

- Seberapa baik kinerja algoritma analisis overlay paralel yang diusulkan dalam penelitian ini dibandingkan dengan penggunaan langsung paradigma komputasi percikan?
- Seberapa besar pengaruh perbedaan kompleksitas poligon geografis terhadap analisis overlay paralel?

(2) Adegan 2: Bandingkan perbedaan performa empat mode: Spark_original, Spark_MBR, Spark_MBR_Hilbert, dan Spark_MBR_Hilbert_R-tree.
Sepuluh kelompok poligon dengan nomor berbeda digunakan untuk analisis overlay dalam tiga mode. Kami akan mencatat waktu penyelesaian proses analisis overlay dan menggambar kurva konsumsi waktu. Selain mempertimbangkan pengaruh perbedaan kompleksitas bentuk poligon geografis, tiga perbaikan penting digunakan dalam aliran algoritma kami: (1) pemfilteran MBR, (2) partisi Hilbert, (3) pembentukan R-tree. Skenario eksperimental ini dapat menjawab: Seberapa besar pengaruh ketiga perbaikan di atas terhadap efisiensi komputasi paralel?

(3) Adegan 3: Pengujian kinerja akselerasi cluster dari algoritma yang diusulkan.
Data eksperimen ditetapkan pada 10 juta poligon geografis. Satu hingga enam server digunakan untuk melakukan analisis overlay dan mencatat perubahan konsumsi waktu dari algoritma analisis overlay dalam penelitian ini. Dalam skenario eksperimental ini, kita dapat melihat rasio akselerasi dan efisiensi paralel dari algoritma yang diusulkan di cluster Spark.

3.7 PERBANDINGKAN PERBEDAAN KINERJA EMPAT MODE

Mode komputasi paralel menggunakan enam node komputasi untuk menghitung aliran sebelum dan sesudah optimasi. Data eksperimen dikumpulkan dari 50.000, 100.000, 250.000, 500.000, 1 juta, 2 juta, 4 juta, 6 juta, 8 juta, dan 10 juta kumpulan data yang tercatat. Statistik konsumsi waktu dari mode komputasi yang berbeda diilustrasikan pada Gambar 3.8.

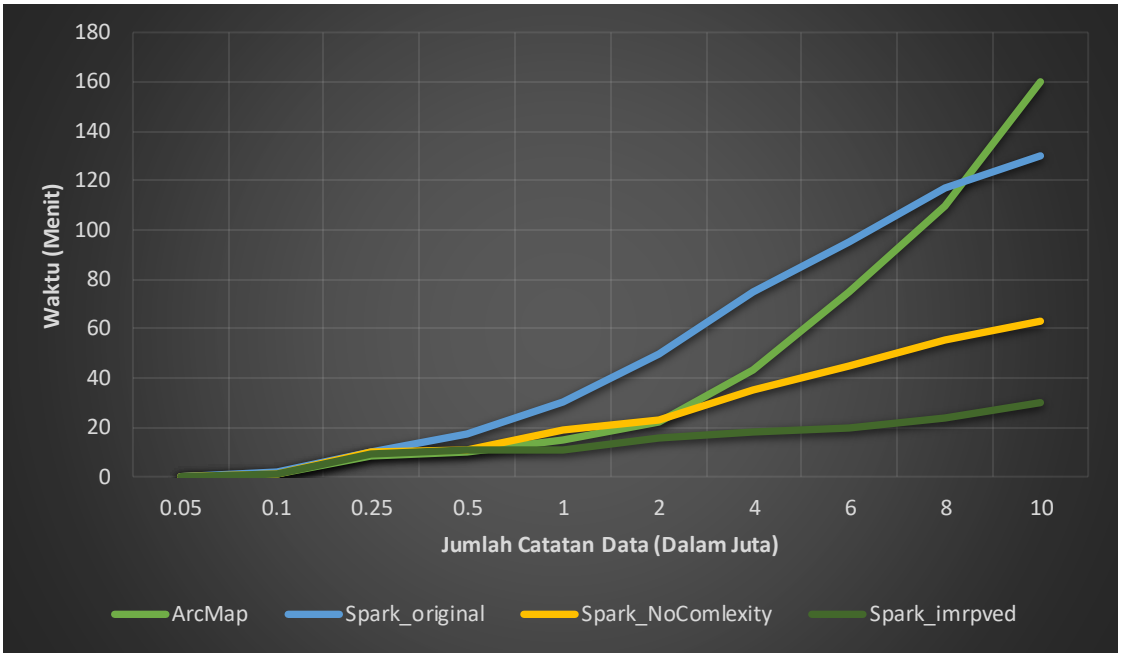
Seperti yang ditunjukkan pada Gambar 3.9, warna biru, merah, kuning, dan abu-abu mewakili waktu yang digunakan oleh mode ArcMap, Spark_original, Spark_NoComlexity, dan Spark_improved. Dengan bertambahnya volume data, keempat kurva konsumsi waktu menunjukkan tren yang meningkat. Perubahan kurva ini menjawab tiga pertanyaan terkait desain skenario eksperimental.

(1) Ketika jumlah poligon kurang dari 10 juta, efisiensi mode Spark_original bahkan lebih rendah dibandingkan mode ArcMap. Ketika jumlah poligon lebih dari 50.000, konsumsi waktu mode Spark_improved lebih sedikit dibandingkan mode ArcMap. Ketika jumlah poligon melebihi 1 juta, mode ArcMap menghabiskan waktu dua kali lebih banyak dibandingkan mode Spark_improved. Seiring bertambahnya jumlah data, konsumsi waktu mode ArcMap meningkat secara dramatis, dan kurva konsumsi waktu mode Spark_improved masih relatif datar.

(2) Efisiensi mode Spark_original lebih rendah dibandingkan mode Spark_improved, dan semakin banyak poligon, semakin jelas tampilannya. Hal ini menunjukkan bahwa efisiensi analisis overlay menggunakan Spark secara langsung sangat rendah, dan

optimasi algoritma harus dilakukan sesuai dengan karakteristik data spasial dan perhitungan geografis.

(3) Dengan membandingkan kurva konsumsi waktu, Spark_improved membutuhkan waktu hampir separuh waktu Spark_NoComlexity, dan ini lebih baik dari yang saya kira. Saya pikir ini mungkin terkait dengan data eksperimen, saya menemukan bahwa ada banyak poligon dengan kompleksitas bentuk yang tinggi dalam data eksperimen. Mungkin banyak poligon besar yang dipartisi ke dalam partisi komputasi yang sama, sehingga menyebabkan data menjadi miring.



Gambar 3.8. Grafik statistik konsumsi waktu dari mode komputasi yang berbeda.

	0.05	0.1	0.25	0.5	1	2	4	6	8	10
Spark_Original	2.6	4	12	20	36	66	119	166	230	300
Spark_MBR	1.9	2	6	8.9	14	22	46	62	80	108
Spark_MBR_Hilbert	1.3	1	7	6.3	7.5	9.9	15	27	36	47
Spark_MBR_Hilbert_R-tree	1.7	1	6	8.2	8.5	11	17	23	32	38

Gambar 3.9. Perbandingan konsumsi waktu dari berbagai strategi optimasi.

Seperti yang ditunjukkan pada Gambar 3.9:

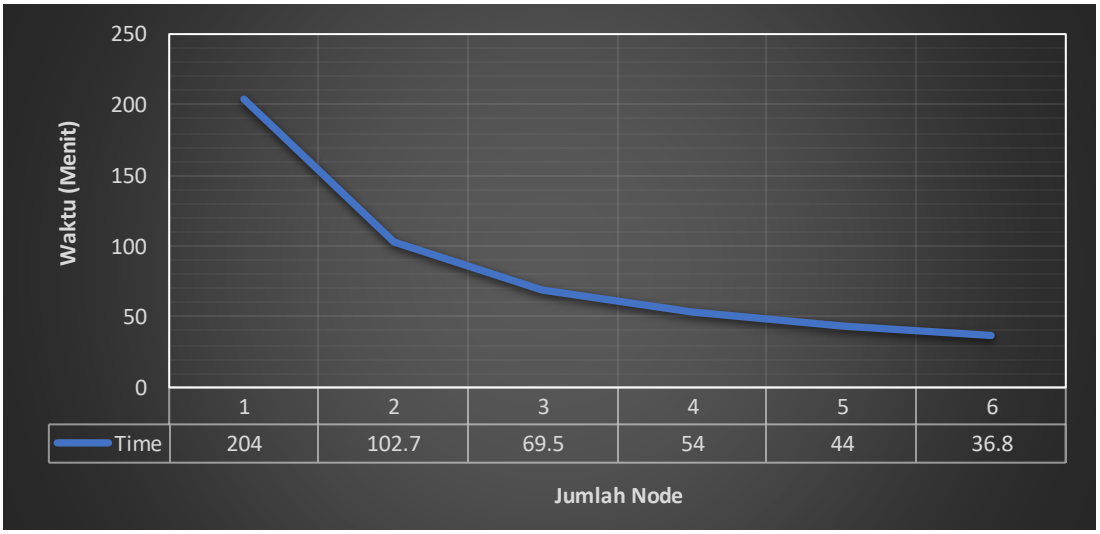
(1) Setelah hanya mengadopsi strategi pemfilteran MBR, efisiensi komputasi overlay meningkat dua hingga empat kali lipat. Oleh karena itu, strategi ini menyaring sejumlah besar perhitungan overlay yang tidak valid. Peningkatan efisiensi spesifik berkaitan dengan ukuran, bentuk, dan distribusi spasial poligon pada lapisan target dan lapisan yang terpotong.

(2) Algoritme partisi Hilbert berdasarkan kompleksitas grafik poligon digunakan untuk mengalokasikan data dari setiap node komputasi. Ketika jumlah data mencapai jutaan, kinerja komputasi bisa berlipat ganda. Seiring bertambahnya jumlah data, keunggulan kinerja komputasi menjadi lebih jelas. Data eksperimen memverifikasi bahwa karakteristik agregasi spasial dari partisi Hilbert yang mempertimbangkan kompleksitas poligon dapat meningkatkan algoritma analisis spasial.

(3) Konstruksi indeks secara umum dapat meningkatkan efisiensi akses data, namun konstruksi indeks itu sendiri dapat mengakibatkan sejumlah overhead komputasi. Setelah menambahkan strategi indeks R-tree berdasarkan dua langkah pertama, waktu penghitungan overlay setiap urutan besarnya sedikit meningkat ketika jumlah data kurang dari 5 juta. Ketika jumlah data melebihi 5 juta, waktu penghitungan overlay berkurang dibandingkan dengan kasus tanpa indeks R-tree. Oleh karena itu, waktu akses data yang dihemat setelah indeks R-tree dibuat mengimbangi waktu yang digunakan oleh indeks itu sendiri.

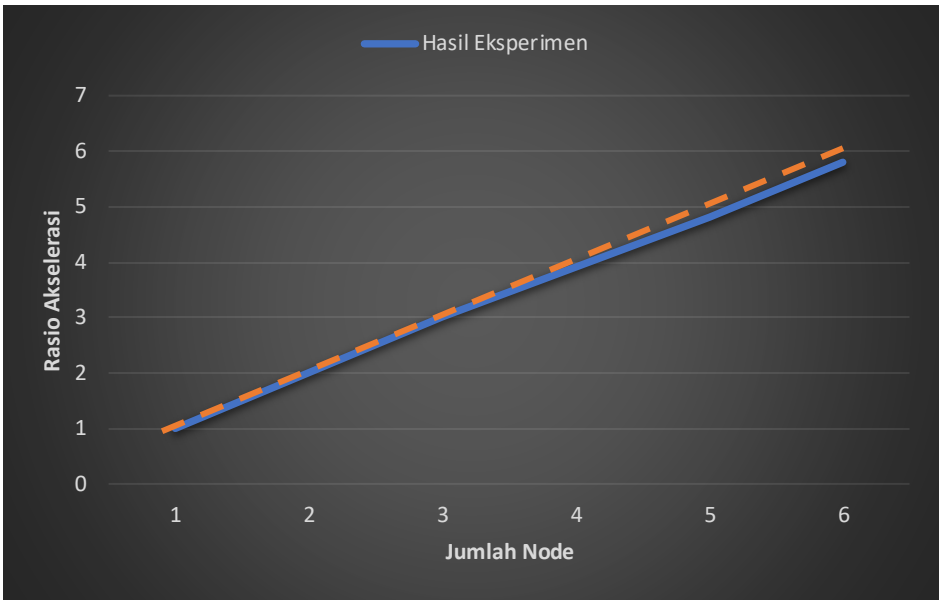
Pengujian Kinerja Akselerasi Cluster dari Algoritma yang Diusulkan

Data eksperimen disatukan menggunakan 10 juta poligon, dan kemudian satu server ditambahkan pada satu waktu. Seiring bertambahnya jumlah server, waktu yang dikonsumsi dalam komputasi paralel berkurang secara signifikan (Gambar 3.10). Namun, tren waktu berjalan menurun seiring dengan bertambahnya jumlah node.

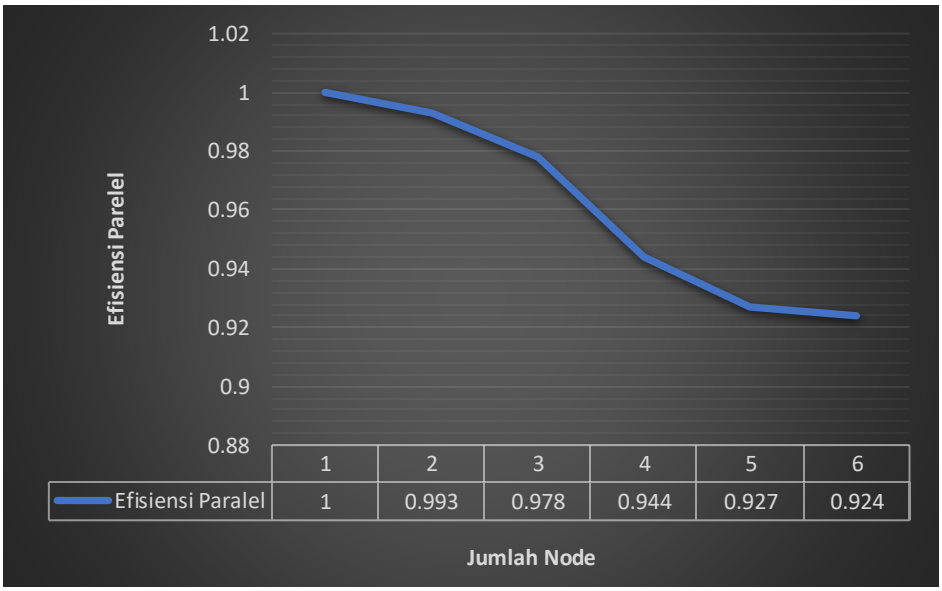


Gambar 3.10. Rata-rata waktu berjalan dari jumlah node yang berbeda.

Seperti yang ditunjukkan pada Gambar 3.11, seiring bertambahnya jumlah server, rasio akselerasi sedikit menurun tetapi hampir linier. Selain itu, Gambar 3.12 mengilustrasikan bahwa dengan bertambahnya jumlah server, efisiensi paralel secara bertahap menurun, dan akhirnya stabil hingga lebih dari 90%. Ini adalah hasil yang baik mengingat peningkatan jumlah server akan meningkatkan sinkronisasi sistem dan overhead jaringan.



Gambar 3.11. Rasio percepatan dari jumlah node yang berbeda.



Gambar 3.12. Efisiensi paralel dari jumlah node yang berbeda.

3.8 ANALISIS PENGUJIAN

Gambar 3.9 menunjukkan bahwa satu komputer dengan ArcMap Soft mencapai efisiensi tinggi dalam analisis overlay volume data kecil dengan mengadopsi algoritma yang masuk akal dan teknologi pemrosesan multithreading yang sangat baik. Selain itu, SDD juga memegang peranan penting. Namun performa ArcMap menurun tajam ketika jumlah record

data mencapai jutaan. Komputasi paralel terdistribusi Spark dapat memecahkan masalah tersebut secara efektif, namun transplantasi sederhana algoritma analisis overlay ke dalam kerangka Spark bukanlah solusi yang masuk akal. Data geografis sebenarnya sering kali tidak terdistribusi secara merata, dan kompleksitas grafik poligon sangat bervariasi, yang akan menyebabkan ketidakseimbangan data yang serius, dan akan berdampak serius pada kinerja komputasi paralel. Ketika volume data mencapai puluhan juta, kinerja algoritma kami meningkat lebih dari 10 kali lipat melalui partisi Hilbert berdasarkan kompleksitas grafis poligon dan indeks R-tree. Selain itu, keunggulan kinerja menjadi lebih nyata seiring dengan peningkatan volume data.

Ketika jumlah data konstan, konsumsi waktu analisis overlay paralel berkurang seiring bertambahnya jumlah server. Namun, tren penurunan waktu berjalan menurun seiring dengan bertambahnya jumlah node. Gambar 3.11 menunjukkan bahwa rasio percepatan hampir linier. Gambar 3.12 mengilustrasikan bahwa efisiensi paralel masih di atas 90% dan tetap stabil ketika jumlah server bertambah menjadi enam, yang berarti efisiensi komputasi yang lebih tinggi dapat dipertahankan ketika cluster komputasi berkembang. Oleh karena itu, ini merupakan metode yang efektif untuk menambahkan node fisik dalam analisis overlay data besar-besaran.

Selain itu, algoritma analisis overlay yang diusulkan juga memiliki beberapa masalah yang perlu diperbaiki, seperti: (1) Poligon besar akan menjangkau beberapa partisi data, yang akan menyebabkan poligon berpartisipasi berulang kali dalam analisis overlay di beberapa server. (2) Dalam proses algoritma saat ini, indeks R-tree dibuat sementara, yang mengarah pada pembuatan indeks berulang kali untuk setiap analisis overlay.

3.8 RANGKUMAN

Dalam analisis overlay paralel performa tinggi, perbedaan kompleksitas bentuk poligon dapat menyebabkan ketidakseimbangan data yang serius. Dalam bab ini, kami mengukur kompleksitas bentuk poligon dari perspektif komputasi geografis dan merancang algoritma analisis overlay paralel berkinerja tinggi dengan mempertimbangkan kompleksitas bentuk poligon. Analisis algoritma menunjukkan bahwa algoritma mengurangi perhitungan overlay yang tidak valid dengan pemfilteran MBR, mencapai penyeimbangan beban dengan menggunakan partisi Hilbert berdasarkan kompleksitas bentuk poligon, dan meningkatkan kecepatan akses data menggunakan indeks R-tree. Eksperimen menunjukkan bahwa ini adalah metode berkinerja tinggi dan dapat mempertahankan kecepatan tinggi dan efisiensi paralel dalam perluasan cluster komputasi.

Dalam bab selanjutnya, kita akan mempelajari dampak distribusi grafis spasial, penyimpanan data spasial, dan metode pengindeksan terhadap efisiensi analisis overlay. Kami juga akan mengoptimalkan penyimpanan indeks spasial melalui teknologi database memori terdistribusi untuk lebih meningkatkan efisiensi analisis overlay paralel.

BAB 4
MODEL MARKOV AUTOMATA SELULER PARALEL

Model Cellular Automata Markov menggabungkan kemampuan model Cellular Automata (CA) untuk mensimulasikan variasi spasial sistem yang kompleks dan prediksi jangka panjang model Markov. Dalam penelitian ini, kami merancang model CA-Markov paralel berdasarkan kerangka MapReduce. Model ini dibagi menjadi dua bagian utama: Model Markov paralel berdasarkan MapReduce (Cloud-Markov), dan metode evaluasi komprehensif perubahan penggunaan lahan berdasarkan automata seluler dan MapReduce (Cloud-CELUC). Memilih Hangzhou sebagai wilayah studi dan menggunakan citra penginderaan jauh Landsat dari tahun 2006 dan 2013 sebagai data eksperimen, kami melakukan tiga eksperimen untuk mengevaluasi model paralel CA-Markov di lingkungan Hadoop. Evaluasi efisiensi dilakukan untuk membandingkan Cloud-Markov dan Cloud-CELUC dengan jumlah data yang berbeda. Hasil penelitian menunjukkan bahwa rasio Cloud-Markov dan Cloud-CELUC mengalami percepatan 3,43 dan 1,86, masing-masing, dibandingkan dengan algoritma serialnya. Uji validitas algoritma prediksi dilakukan dengan menggunakan model paralel CA-Markov untuk mensimulasikan perubahan penggunaan lahan di Hangzhou pada tahun 2013 dan menganalisis hubungan antara hasil simulasi dan hasil interpretasi citra penginderaan jauh. Koefisien Kappa untuk lahan konstruksi, lahan cagar alam, dan lahan pertanian masing-masing adalah 0,86, 0,68, dan 0,66, yang menunjukkan validitas model paralel. Perubahan penggunaan lahan Hangzhou pada tahun 2020 diprediksi dan dianalisis. Hasilnya menunjukkan bahwa area sentral lahan konstruksi meningkat pesat karena sistem transportasi yang berkembang dan sebagian besar dialihkan dari lahan pertanian.

4.1 PENDAHULUAN

Mempelajari perubahan penggunaan/tutupan lahan pada waktu dan tempat yang berbeda serta memperkirakan struktur penggunaan lahan dan tata ruang dapat memberikan dukungan ilmiah terhadap pemanfaatan sumber daya lahan regional, perlindungan lingkungan ekologi regional, dan pembangunan sosial dan ekonomi berkelanjutan.

Banyak peneliti telah mengusulkan model simulasi dan prediksi perubahan penggunaan lahan mereka sendiri, seperti CLUE, CLUE-S, Cellular Automata (CA), Markov Chain, SLEUTH, dan model logistik spasial. Sejak awal tahun 1980an ketika Wolfram pertama kali mengusulkan model CA, banyak penelitian telah dilakukan untuk menggunakan model CA untuk mensimulasikan perubahan penggunaan lahan perkotaan dan para peneliti telah mengintegrasikan metode atau model lain, seperti jaringan saraf, mesin vektor dukungan (SVM), optimasi ant-koloni, dan rantai Markov, ke dalam model CA untuk mensimulasikan dan memantau perubahan penggunaan lahan.

Model CA-Markov adalah salah satu model CA diperluas yang paling banyak digunakan dan digunakan dalam prediksi dan simulasi perubahan penggunaan lahan di banyak negara, seperti Amerika Serikat, Brasil, Portugal , Mesir, Ethiopia, Bangladesh, Malaysia, dan Cina. Hal

ini juga telah diterapkan pada penelitian evolusi pola pemukiman perkotaan, proses perubahan vegetasi dinamis spasial, dan pengalihan lahan melintasi wilayah metropolitan.

Karena simulasi dan prediksi perubahan penggunaan lahan melibatkan sejumlah besar data dan perhitungan, dalam beberapa tahun terakhir, beberapa penelitian telah merancang algoritma CA paralel pada komputasi paralel Central Processing Unit (CPU), Message Passing Interface (MPI), Graphics Processing Unit (GPU) parallel, dan GPU/CPU hybrid parallel untuk mensimulasikan pertumbuhan perkotaan. Namun metode CA paralel tidak dapat menangani hubungan antar partisi setelah suatu area penelitian dibagi menjadi beberapa bagian sehingga menghasilkan hasil prediksi akhir yang berbeda-beda, sedangkan metode Markov tradisional mampu menjaga keutuhan seluruh area penelitian namun menghasilkan kurangnya hubungan spasial sel tanah. Di sisi lain, dengan berkembangnya teknologi dan aplikasi Big Data, MapReduce adalah metode yang menjanjikan untuk meningkatkan efisiensi berjalan algoritma serial tradisional dan telah diterapkan dan terbukti efektif dalam banyak kasus. Rathore dkk. mengusulkan aplikasi pemrosesan dan analisis gambar penginderaan jauh secara real-time. Raojun dkk. mengusulkan algoritma prediksi tautan paralel berdasarkan MapReduce. Wiley K.dkk. menganalisis grafik astronomi berdasarkan MapReduce, sementara Almeer menggunakan Hadoop untuk menganalisis gambar penginderaan jauh, meningkatkan efisiensi membaca dan menulis secara batch. Untuk model CA Markov, kerangka MapReduce tidak hanya mampu melakukan pemrosesan paralel secara efisien, namun juga dapat menjadi penghubung dengan model CA-Markov: “Map” berhubungan dengan proses CA untuk mewujudkan paralelisme unit penggunaan lahan -perubahan prediksi; “Pengurangan” mengacu pada proses Markov untuk mencapai prediksi perubahan penggunaan lahan secara keseluruhan. Namun, karena masalah utama segmentasi dan koneksi masih belum terselesaikan, hanya ada sedikit penelitian mengenai model paralel CA-Markov untuk prediksi perubahan penggunaan lahan melalui kerangka MapReduce.

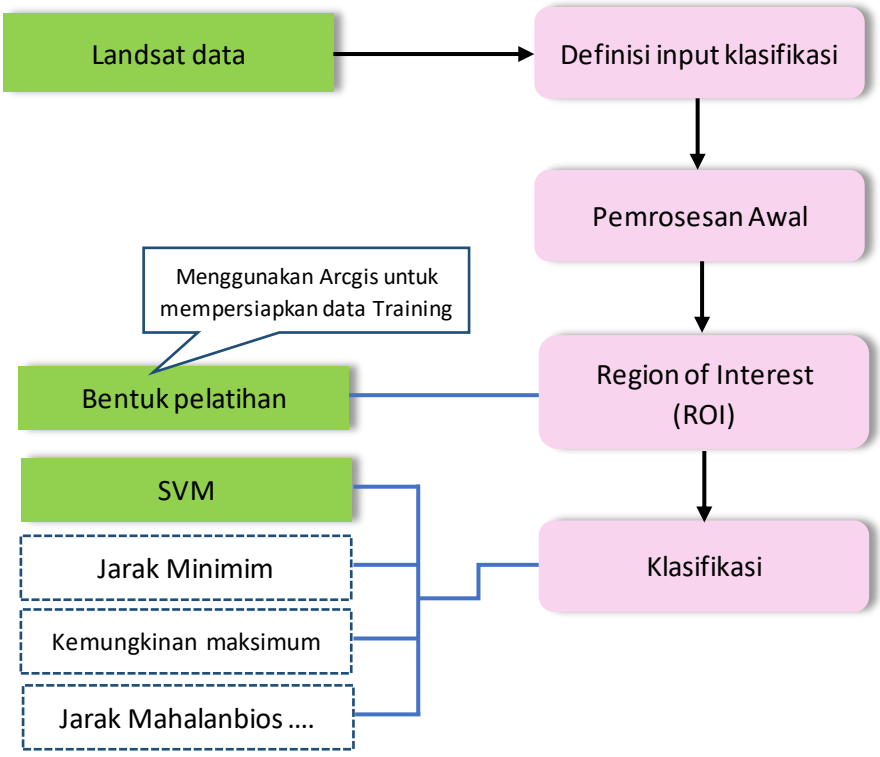
Berdasarkan analisis mendalam terhadap paralelisme model CA Markov, bab ini pertama-tama mengusulkan solusi paralel yang menggunakan kerangka kerja MapReduce untuk menyempurnakan model CA Markov dalam prediksi perubahan penggunaan lahan. Model CA-Markov paralel tidak hanya dapat menyelesaikan kontradiksi bahwa model CA-Markov tradisional tidak dapat secara bersamaan mewujudkan integritas dan segmentasi untuk simulasi dan prediksi perubahan penggunaan lahan, namun juga dapat memastikan efisiensi dan akurasi serta mewujudkan prediksi perubahan penggunaan lahan. dalam lingkungan komputasi awan.

4.2. WILAYAH EXPERIMEN DAN DATA

Hangzhou terletak di pantai tenggara Tiongkok, yang merupakan pusat politik, ekonomi, budaya, dan keuangan Provinsi Zhejiang. Hangzhou memiliki topografi yang kompleks: Bagian barat merupakan daerah perbukitan dengan pegunungan utama, termasuk Gunung Tianmu, dan bagian timur merupakan daerah dataran dengan dataran rendah dan jaringan sungai yang padat.

Citra penginderaan jauh Landsat TM 2006 dan Landsat8 2013 dengan resolusi 30 m di area penelitian diunduh dari http://www.gscloud.cn/. Kumpulan data eksperimen lainnya mencakup DEM dengan resolusi 30 m, data jaringan jalan, data lokasi lalu lintas, dan data alamat lokasi.

Saat ini, banyak penelitian yang mengembangkan metode identifikasi citra otomatis untuk mengklasifikasikan citra resolusi tinggi, khususnya citra kendaraan udara tak berawak (UAV). Karena citra Landsat merupakan citra beresolusi sedang, kami memilih menggunakan metode semi-manual untuk melakukan praproses dan interpretasi guna mendapatkan data penggunaan lahan menggunakan ENVI 5.3 dan ArcGIS 10.2. Seperti ditunjukkan pada Gambar 4.1, alur kerja klasifikasi citra Landsat mencakup empat langkah utama, yaitu definisi masukan klasifikasi, prapemrosesan, wilayah yang diminati, dan klasifikasi. Langkah-langkah pemrosesan terutama mencakup koreksi geo, rektifikasi geometri, atau registrasi gambar, kalibrasi radiometrik dan koreksi atmosfer, dan koreksi topografi. Pengklasifikasi mesin vektor pendukung (SVM) dipilih untuk klasifikasi penggunaan lahan.



Gambar 4.1. Alur klasifikasi citra Landsat.

Secara umum tipe penggunaan lahan meliputi lahan budidaya, hutan, padang rumput, perairan, lahan konstruksi, kebun, lahan transportasi, lahan tidak terpakai, dan lahan rawa. Dalam percobaan kami, empat tipe penggunaan lahan didefinisikan dalam bentuk pelatihan setelah diklasifikasi ulang. Metode reklasifikasi tipe penggunaan lahan pada lahan konstruksi (B), lahan pertanian (A), cagar alam (N), dan wilayah perairan (W) didefinisikan seperti ditunjukkan pada Tabel 4.1.

Tabel 4.1. Reklasifikasi penggunaan lahan.

Level 1	Level 2	Definisi
Lahan konstruksi (B)	Tanah untuk konstruksi (B1), tanah untuk transportasi (B2)	Tanah untuk bangunan dan struktur.
Lahan pertanian (A)	Lahan budidaya (A1), kebun (A2)	Lahan produksi pertanian.
Luas perairan (W)	Perairan (W1), rawa (W2)	Permukaan sungai, permukaan danau, rawa.
Cagar Alam (N)	Hutan (N1), padang rumput (N2), lahan tak terpakai (N3)	Lahan dengan sedikit atau tanpa aktivitas manusia itu tidak termasuk lahan pertanian,lahan konstruksi, dan perairan.

4.3 MAPREDUCE

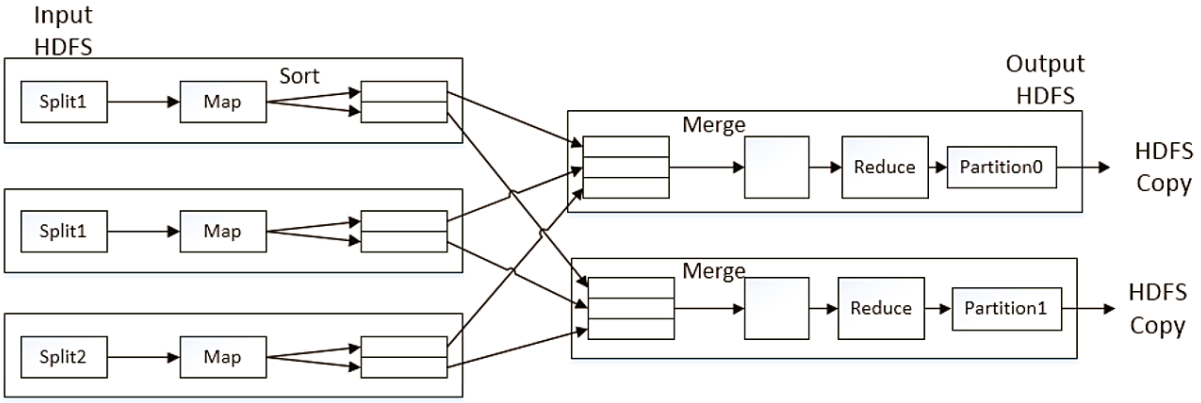
Program MapReduce terdiri dari dua fungsi, Map dan Reduce. Kedua fungsi ini mengambil kunci/nilai (pasangan kunci-nilai) sebagai input dan output. Fungsi Map menerima pasangan nilai kunci yang dimasukkan pengguna (k_1, v_1) dan memprosesnya untuk menghasilkan pasangan nilai kunci (k_2, v_2) sebagai hasil antara. Kemudian, nilai-nilai terkait dari semua kunci perantara yang sama (k_2) dikumpulkan untuk menghasilkan daftar nilai untuk $k_2, list\,(v_2)$, yang digunakan sebagai input ke fungsi Reduce dan diproses oleh fungsi Reduce untuk mendapatkan daftar hasil akhir (k_3, v_3). Prosesnya dapat dinyatakan dengan rumus berikut:

Persamaan 1 (Map), Persamaan 2 (Reduce):

$$\text{Map: } (k_1, v_1) - \text{list}(k_2, v_2)$$

$$\text{Reduce: } (k_2, list\,(v_2)) - \text{list}(k_3, v_3)$$

Kerangka kerja MapReduce ditunjukkan pada Gambar 2:



Gambar 4.2. Ikhtisar kerangka proses MapReduce.

4.4 MODEL CA MARKOV

Model Markov didasarkan pada teori proses acak. Dalam model ini, dengan mempertimbangkan probabilitas keadaan awal dan transisi keadaan, hasil simulasi tidak ada hubungannya dengan kondisi historis sebelum kondisi saat ini, yang dapat digunakan untuk menggambarkan perubahan penggunaan lahan dari satu periode ke periode lainnya. Kita juga

dapat menggunakan ini sebagai dasar untuk memprediksi perubahan di masa depan. Perubahan ditemukan dengan membuat matriks probabilitas transisi perubahan penggunaan lahan dari periode t hingga $t+1$, yang merupakan dasar untuk memprediksi perubahan penggunaan lahan di masa depan. (Persamaan 3)

$$S(t+1) = P_{ij} \times S(t)$$

$S(t+1)$ menunjukkan keadaan sistem penggunaan lahan masing-masing pada waktu $t+1$ dan t. P_{ij}adalah matriks transisi keadaan.

CA memiliki empat komponen dasar: sel dan ruang sel, keadaan sel, lingkungan, dan aturan transisi. Model CA dapat diungkapkan sebagai berikut: (Persamaan 4)

$$S(t+1) = f(S(t), N)$$

Dalam rumusnya, S adalah himpunan keadaan dengan keadaan berhingga dan diskrit, t dan $t+1$adalah momen berbeda, N adalah lingkungan sel, dan f adalah aturan transisi sel ruang lokal.

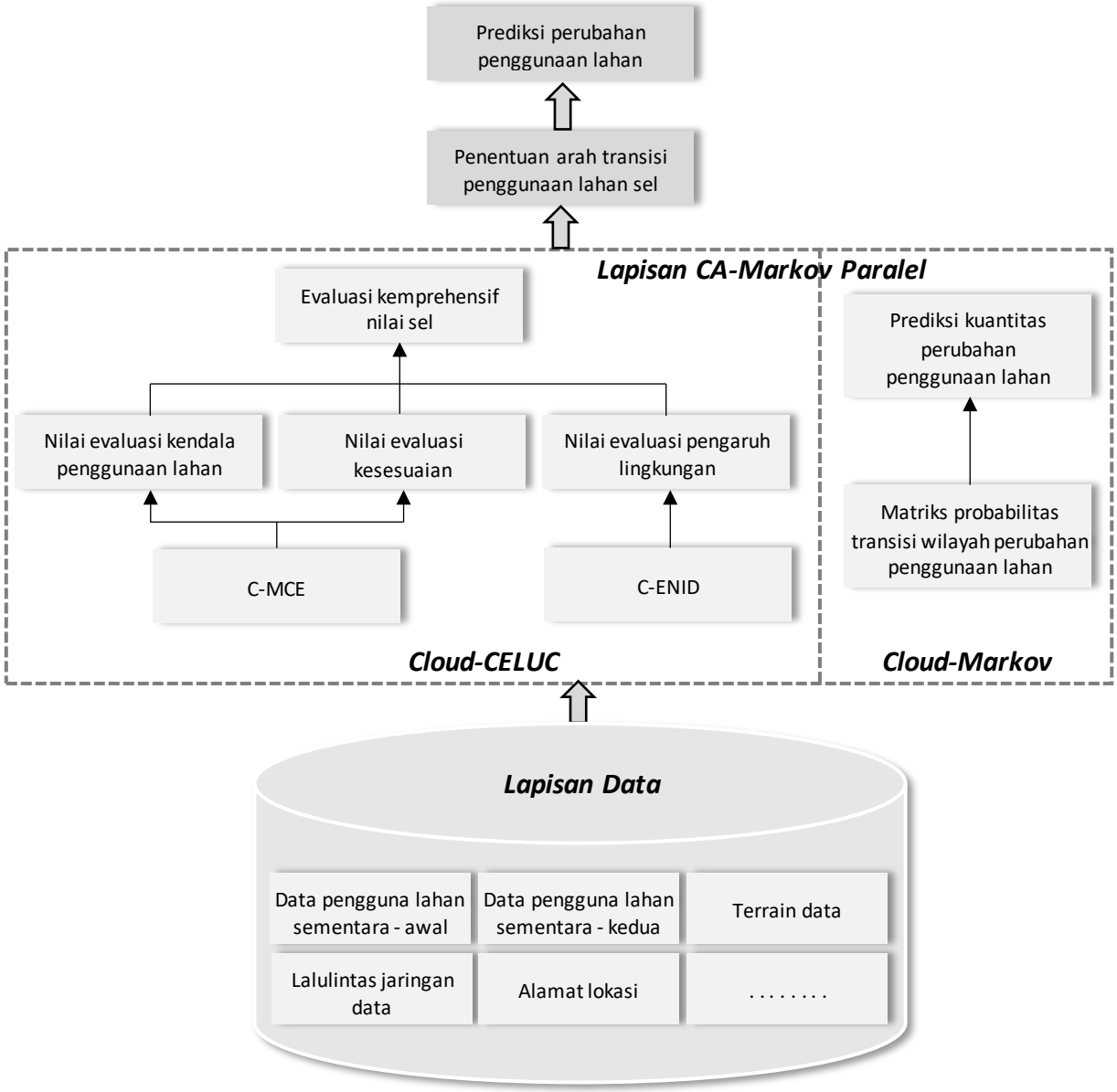
Biasanya, untuk membuat automata seluler mensimulasikan lingkungan nyata dengan lebih baik, variabel batasan ruang β perlu diperkenalkan untuk mengekspresikan medan topografi, serta batasan adaptif dan batasan restriktif dari faktor pengaruh spasial pada sel. Rumusnya menjadi (Persamaan 5)

$$S(t+1) = f(S(t), N, \beta)$$

Model Markov yang terpisah kurang memiliki pengetahuan spasial dan tidak mempertimbangkan distribusi spasial faktor geografis dan tipe penggunaan lahan, sedangkan model CA-Markov menambahkan fitur spasial ke model Markov, menggunakan filter automata seluler untuk menciptakan faktor bobot dengan karakter spasial , dan mengubah keadaan sel sesuai dengan keadaan sel yang berdekatan dan aturan transisi.

Struktur CA-Markov Paralel

Gambar 4.3 adalah struktur paralel CA-Markov melalui kerangka MapReduce. Strukturnya berisi empat lapisan dari bawah ke atas: Lapisan data, lapisan model CA-Markov paralel, lapisan penentuan arah transisi penggunaan lahan, dan lapisan prediksi penggunaan lahan.



Gambar 4.3. Struktur automata seluler paralel Markov melalui MapReduce.

Data percobaan mencakup gambar penginderaan jauh dua fase, data medan, data jaringan lalu lintas, dan data alamat lokasi. Model paralel CA-Markov pada dasarnya dapat dibagi menjadi dua bagian: Algoritme Markov paralel berdasarkan MapReduce (Cloud-Markov), dan metode evaluasi komprehensif perubahan penggunaan lahan berdasarkan MapReduce (Cloud-CELUC). Cloud-Markov digunakan untuk menghitung matriks probabilitas transisi area perubahan penggunaan lahan.

Cloud-CELUC mencakup tiga bagian utama: Evaluasi pengaruh lingkungan di bawah lingkungan komputasi awan (C-ENID), evaluasi multikriteria dalam lingkungan komputasi awan (C-MCE), dan evaluasi komprehensif penggunaan lahan. Diantaranya, C-ENID adalah model CA paralel melalui MapReduce, dan C-MCE adalah algoritma MapReduce untuk menghitung nilai evaluasi terbatas dan evaluasi kesesuaian.

Alur Kerja CA-Markov Paralel

Seperti ditunjukkan pada Gambar 4, aliran model paralel CA-Markov memiliki lima langkah besar, sebagai berikut:

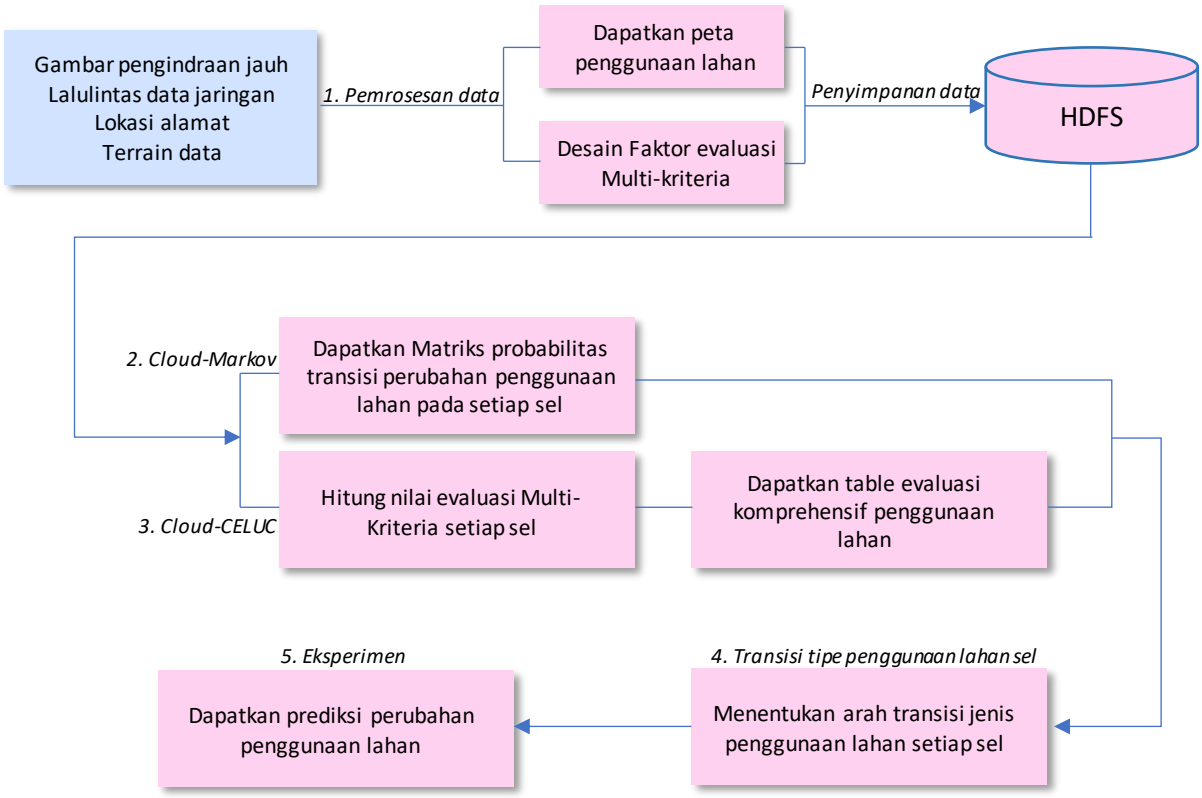
(1) **Pemrosesan data**: Melakukan pra-pemrosesan dan interpretasi citra penginderaan jauh ke peta penggunaan lahan, merancang faktor evaluasi multikriteria, dan menyimpan citra, data penggunaan lahan, dan faktor evaluasi multikriteria ke dalam Hadoop HDFS.

(2) **Markov Paralel:** Menggunakan metode overlay, menganalisis gambar dua fase dan data penggunaan lahan untuk mendapatkan probabilitas transisi tipe penggunaan lahan setiap sel, dan menghitung jumlah total sel di setiap arah transisi tipe penggunaan lahan, dan menghitung matriks probabilitas transisi kawasan untuk setiap tipe penggunaan lahan.

(3) **CA Paralel:** Di Cloud-CELUC, C-ENID digunakan untuk menghitung nilai pengaruh lingkungan sel. C-MCE dirancang untuk menghitung nilai evaluasi multikriteria, termasuk nilai evaluasi kendala dan nilai evaluasi kesesuaian. Nilai-nilai ini kemudian digunakan untuk menghitung tabel statistik evaluasi komprehensif penggunaan lahan.

(4) **Tahap penentuan arah transisi:** Pembacaan loop probabilitas transisi setiap sel dari tabel statistik nilai evaluasi komprehensif pada tahap CA paralel dan menggabungkan matriks probabilitas transfer area setiap tipe penggunaan lahan pada tahap Markov paralel untuk memutuskan arah transisi tipe penggunaan lahan suatu sel.

(5) **Prediksi perubahan penggunaan lahan:** Dalam percobaan kami, kami menggunakan data dari tahun 2006 untuk memprediksi perubahan penggunaan lahan tahun 2013 dan kemudian mengevaluasi ketepatan model paralel CA-Markov dengan koefisien Kappa. Prediksi perubahan penggunaan lahan untuk tahun 2020 kemudian diperoleh.



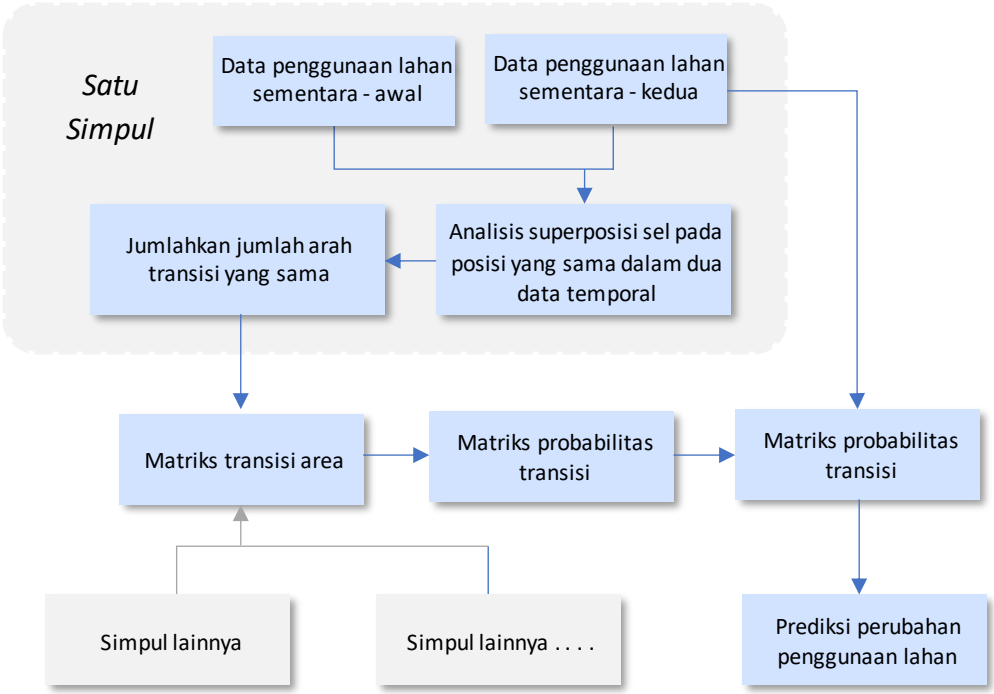
Gambar 4.4. Aliran Paralel Cellular Automata-Markov (CA-Markov).

Pemrosesan Paralel Model Markov (Cloud-Markov)

Dengan menggunakan metode overlay untuk menganalisis gambar raster dua fase dan data penggunaan lahan dengan posisi spasial yang sama, diperoleh arah transisi setiap sel dan menghitung jumlah sel pada setiap arah transisi, kemudian diperoleh matriks perpindahan luas setiap lahan. Jenis penggunaan dan menghitung matriks probabilitas luas penggunaan lahan. Matriks perpindahan wilayah setiap tahun kemudian diperoleh dengan menggunakan matriks probabilitas dibagi dengan interval kedua gambar raster tersebut. Kemudian, fungsi MapReduce dari model Markov yang diberikan di wilayah penelitian adalah sebagai berikut: (Persamaan 6=Map; Persamaan 7=Combiner; Persamaan 8=Reduce)

$$\text{Map}: (N, (T_1, T_2)) \rightarrow \ list(C_{mk}, i)$$
$$\text{Combiner}: M, (C_{mk}, List(i)) \rightarrow \ list(C_{mk}, s)$$
$$\text{Reduce}: L, (C_{mk}, List(i)) \rightarrow \ list(C_{mk}, s)$$

dimana N adalah offset baris dari baris input, T_1 mewakili semua jenis penggunaan lahan pada gambar raster sebelumnya, T_2 mewakili semua jenis penggunaan lahan pada gambar raster selanjutnya, C_{mk} mewakili konversi sel dari penggunaan lahan M ke tipe penggunaan lahan k, i menunjukkan nomor transfer sel C_{mk}, M adalah jumlah tipe penggunaan lahan di node MapReduce, s adalah jumlah total sel C_{mk} di node MapReduce, L adalah jumlah total jenis penggunaan lahan, dan q adalah nilai gabungan jumlah sel total dan probabilitas transisi konversi C_{mk} di seluruh wilayah studi. Misalnya, "100-3,12%" berarti terdapat 100 sel dengan konversi C_{mk}, dan probabilitas transisi adalah 3,12%. Gambar 4.5 merupakan alur algoritma Cloud-Markov. Setelah menjumlahkan jumlah arah transisi tipe penggunaan lahan yang sama di setiap node, matriks transisi area dihitung pada tahap Reduce.



Gambar 4.5. Alur algoritma Cloud-Markov.

Langkah-langkahnya adalah sebagai berikut:

(1) Tahapan peta:
   a. Input <Key,Value>
   b. Analisis konversi tipe penggunaan lahan sel raster
      Pada langkah ini, dengan membandingkan sel pada posisi yang sama antara dua gambar raster ini, kami memperoleh daftar C_{mk}. Jika nilai C_{mk} adalah 'B-A', berarti tipe penggunaan lahan sel dengan posisi yang sama pada gambar raster berbeda diubah dari B menjadi A.
   c. Output <Keyi,Value>
      di mana kuncinya adalah arah konversi C_{mk}, dan Nilai adalah bilangan bulat sama dengan 1.

(2) Tahap penggabung:
   a. Input <key,value>
   b. Hitung jumlah setiap arah konversi tipe penggunaan lahan di setiap node
      <Key,Value> adalah pasangan kunci-nilai (C_{mk},s), dengan C_{mk} adalah arah konversi dan s adalah jumlah total sel dalam node MapReduce tempat terjadinya arah konversi jenis penggunaan lahan C_{mk}.
   c. Output <Key,value>
      Pasangan nilai kunci keluaran (C_{mk}, s).

(3) Kurangi tahap:
   a. Input <Key,value>
   b. Hitung probabilitas transisi
      Rumus perhitungan matriks probabilitas transisi didefinisikan sebagai berikut:
      (Persamaan 9)

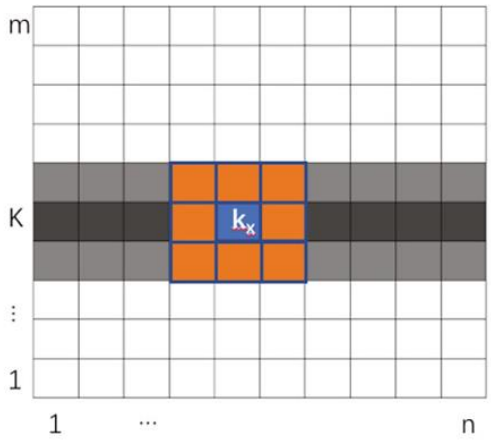
$$P_{mk} = \frac{V_{mk}}{S_m}$$

      dimana V_{mk}adalah nilai penjumlahan C_{mk} dari seluruh node MapReduce dan S_m adalah jumlah sel gambar raster awal yang tipe penggunaan lahannya m.
   c. Output <Key,Value>
      <Key,Value> adalah pasangan kunci-nilai ($V_{mk}P_{mk}$), denganV_{mk} adalah matriks konversi area tipe penggunaan lahan dan P_{mk} adalah matriks probabilitas transisi.

4.5 CLOUD-CELUC

Pemrosesan Lingkungan Sel

Untuk mendapatkan nilai pengaruh lingkungan suatu sel memerlukan pembacaan keadaan sel tetangganya. Metode evaluasi pengaruh lingkungan secara umum meliputi Von Neumann dan Moore. Gambar 4.6 menunjukkan cara membaca lingkungan seluler, di mana kami memilih metode 3 × 3 Moore untuk merancang algoritma kami.



Gambar 4.6. Pembacaan lingkungan seluler.

Gambar 4.7 menunjukkan proses reduksi dimensi seluler, dimana struktur daftar digunakan untuk menyimpan semua sel, sehingga kita dapat membaca sel lingkungan setiap sel melalui indeks baris dan indeks kolom sel. Gambar raster dua dimensi direduksi menjadi array satu dimensi yang dapat mengurangi pertukaran data antar setiap node selama proses MapReduce. Misalnya, sel tetangga sel K_x berasal dari garis $K-1, K$, dan $K+1$ dicatat sebagai $\mathrm{K}-1_{\mathrm{x}}-1, \mathrm{K}-1_{\mathrm{x}}, \mathrm{K}-1_{\mathrm{x}}+1, \mathrm{K}_{\mathrm{x}}-1, \mathrm{K}_{\mathrm{x}}+1, \mathrm{K}-1_{\mathrm{X}}-1, \mathrm{K}+1_{\mathrm{x}}$ dan $\mathrm{K}+1_{\mathrm{x}}+1$, lalu disimpan sebagai struktur array ke dalam HDFS.

Gambar 4.7. Reduksi dimensi seluler.

Faktor Evaluasi Multikriteria

Buku ini menggunakan metode evaluasi multikriteria (MCE) untuk menghitung nilai evaluasi kendala dan evaluasi kesesuaian. Tujuan MCE adalah untuk memilih solusi keputusan optimal dalam solusi terbatas (tak terbatas) yang di antara solusi tersebut terdapat konflik dan hidup berdampingan.

Kriteria evaluasi MCE dibagi menjadi dua bagian: Faktor yang sesuai dan faktor kendala. Faktor yang sesuai digunakan untuk menormalkan faktor yang mempengaruhi ke nilai terukur yang berkelanjutan, dan faktor kendala digunakan untuk mengkategorikan fitur spasial berdasarkan karakteristik spasialnya. Nilai faktornya bertipe Boolean dengan nilai 0 atau 1.

Faktor kesesuaian didefinisikan dan distandarisasi dalam Tabel 4.2, di mana delapan faktor didefinisikan untuk membedakan jarak dari sel ke tujuan umum yang berbeda, dan menurut tipe penggunaan lahan yang berbeda, nilai tertimbang dari masing-masing faktor ditentukan dalam Tabel 4.3. Kemudian, metode proses hierarki analitik (AHP) dan metode penilaian ahli digunakan untuk menghitung bobot faktor kesesuaian, yang ditunjukkan pada Tabel 4.2. Faktor perairan dan gradien didefinisikan sebagai faktor kendala.

Tabel 4.2. Klasifikasi faktor kesesuaian.

Nama Faktor	Definisi	Klasifikasi
FreLev	Jarak dari sel ke jalan raya.	Jarak sel dari jalan utama atau pusat kota: 0-250, 250-500, 500-750, 750-1000, dan 1000-1250 m.
TownLev	Jarak dari sel ke pusat kota.	
SubLev	Jarak dari sel ke stasiun kereta bawah tanah.	Jarak sel ke stasiun kereta bawah tanah atau bus, jalan lain: 0-100, 100-200, 200-300, 300-400, dan 400-500 m.
BusLev	Jarak dari sel ke halte bus.	
MainLev	Jarak dari sel ke jalan lain.	
TraLev	Jarak dari sel ke stasiun kereta.	Jarak sel ke stasiun kereta atau bus: 0-200, 200-400, 400-600, 600-800, dan 800-1000 m.
StaLev	Jarak dari sel ke stasiun bus.	
CityLev	Jarak dari sel ke pusat daerah.	Jarak sel ke jalan utama atau pusat kabupaten: 500-1000, 1000-1500, 1500-2000, dan 2000-2500 m.

Tabel 4.3. Parameter bobot faktor kesesuaian.

Nama Faktor	Tanah Pertanian	Tanah Konstruksi	Cagar Alam
FreLev	0.04885	0.0461	0.0781
TownLev	0.1239	0.1332	0.1010
SubLev	0.0621	0.1320	0.0133
BusLev	0.0721	0.1110	0.0513
MainLev	0.0921	0.1102	0.0749
TraLev	0.0423	0.1333	0.0201
StaLev	0.0623	0.1321	0.0203
StaLev	0.0923	0.2021	0.0103

Nilai faktor kendala adalah Boolean yang ditentukan oleh tipe penggunaan lahan. Penelitian ini mendefinisikan faktor kendala perbukitan dengan kemiringan lebih dari 25 derajat, serta perairan dan cadangan ekologi sama dengan 0, karena tipe penggunaan lahan ini jarang berubah.

Algoritma Cloud-CELUC Paralel

Cloud-CELUC hanya memerlukan fungsi Peta untuk menghitung faktor dan mendapatkan nilai evaluasi yang komprehensif. Fungsi Reduce dari Cloud-CELUC hanya digunakan untuk menampilkan hasilnya. Fungsi Peta didefinisikan sebagai berikut:
(Persamaan 10):

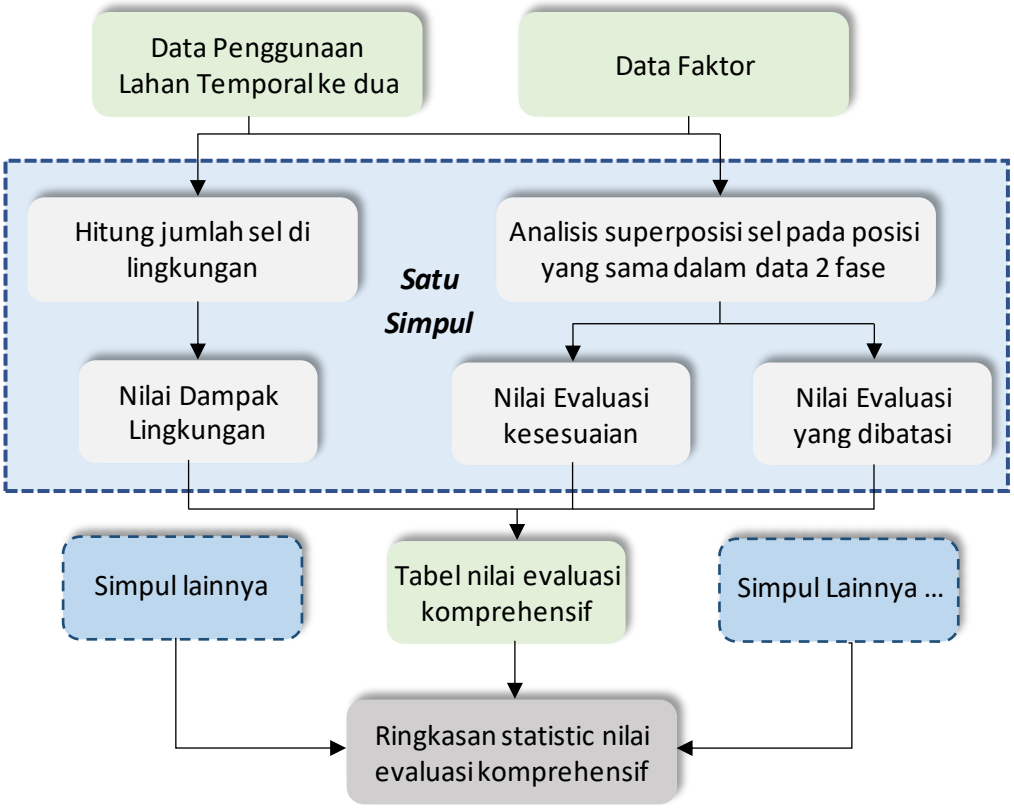
$$\text{Map: } \left(N, (i, H_1, H_2, H_3, H_4, \dots H_m)\right) \rightarrow ((i,j), (L_1, L_2, L_3, \dots L_m))$$

dimana N adalah offset garis dari baris input, i adalah indeks garis dari gambar raster, j adalah indeks kolom dari gambar raster, H_1 adalah nilai keadaan sel yang akan dihitung, dan H_2 dan H_3 menunjukkan uplink dan downlink nilai status sel $H_1.H_4$ ... H_m adalah berbagai faktor kendala dan faktor kesesuaian yang sesuai dengan sel, dan L_m adalah nilai yang digabungkan dengan nilai evaluasi gabungan sel dari arah transisi dan arah transisi sel yang

sesuai. Misalnya, jika L_1 adalah 'ba-1.2234', maka nilai evaluasi sel (i, j) dari tipe penggunaan lahan awal 'b' hingga tipe penggunaan lahan akhir 'a' adalah '1.2234'.

Alur algoritma Cloud-CELUC ditunjukkan pada Gambar 4.8, dimana tabel nilai evaluasi komprehensif diperoleh setelah setiap node dengan menghitung nilai dampak lingkungan, nilai evaluasi kesesuaian, dan nilai evaluasi kendala. Langkah-langkahnya adalah sebagai berikut:

a. Inputan <key,value>
b. Hitung nilai evaluasi pengaruh lingkungan (NID)



Gambar 4.8. Algoritma Cloud-comprehensive-evaluation value (CELUC).

Berdasarkan H_1 dan H_3, tingkat pengaruh lingkungan setiap sel di H_2 dihitung. Rumus perhitungan nilai evaluasi derajat pengaruh lingkungan sel (i, j) yang bersesuaian dengan suatu kelas pada waktu tertentu adalah sebagai berikut: (Persamaan 11)

$$NID_a = \frac{1}{h-1}\sum Yes(S_{ij} == a)$$

Dimana h adalah jumlah sel tetangga, dan $Yes(S_{ij} == a)$ digunakan untuk menilai apakah tipe penggunaan lahan sel tetangga (i, j) adalah a atau bukan. Jika a adalah 1, maka $Yes(S_{ij} == a)$ menghasilkan 1, jika tidak maka menghasilkan 0.

c. Hitung Nilai Evaluasi Kesesuaian (SEV)
Rumus perhitungan nilai evaluasi kesesuaian adalah sebagai berikut: (Persamaan 12)

$$SEV_a = \sum V_{a\delta} \times DIS_{ij\delta}$$

dimana $V_{a\delta}$ adalah bobot faktor δ sesuai dengan tipe penggunaan lahan $'a'$, $DIS_{ij\delta}$ adalah faktor kesesuaian sel (i,j) yang didefinisikan pada Tabel 4.1.

d. Menghitung nilai evaluasi kendala (CEV) Rumus evaluasi kendala adalah sebagai berikut: (Persamaan 13)

$$CEV_a = \prod Yes(K_{ij})$$

dimana $Yes(K_{ij})$ mewakili nilai evaluasi batasan sel (i,j) yang sesuai dengan faktor batasan $'k'$. Jika sel dibatasi, $Yes(K_{ij})$ mengembalikan 0, jika tidak, ia mengembalikan 1.

e. Hitung nilai evaluasi komprehensif (CELUC)
Berdasarkan ketiga nilai evaluasi di atas, yaitu $NID, SEV, \text{dan } CEV$, maka dilakukan evaluasi komprehensif. Rumusnya didefinisikan sebagai berikut: (Persamaan 14)

$$CELUC_a = NID_a \times SEV_a \times CEV_a$$

dimana a adalah tipe penggunaan lahan.

f. Output <Key,Value>
<Key,Value> adalah pasangan kunci-nilai $((i,j), CELUC)$ dengan i adalah indeks baris sel, j adalah indeks kolom sel, dan $CELUC$ adalah nilai evaluasi komprehensif sel.

4.6 KONVERSI JENIS PENGGUNAAN LAHAN SEL

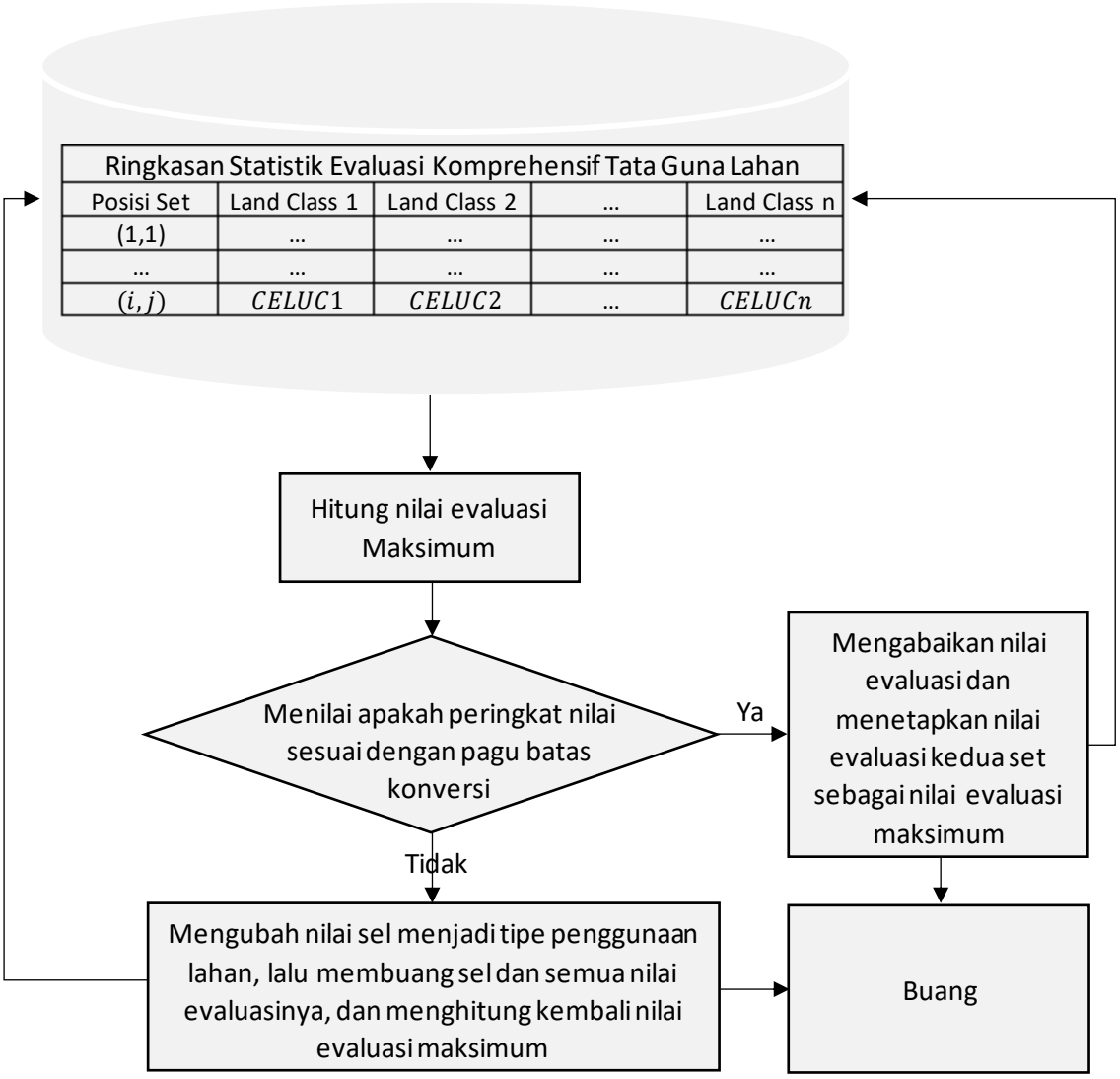
Metode kompetisi penggunaan lahan multi-tujuan digunakan untuk mencapai konversi tipe penggunaan lahan sel, yang memecahkan masalah ketika sel mengalami konflik dalam konversi tipe penggunaan lahan. Misalnya, jika terdapat N tipe penggunaan lahan, sel (i,j) mungkin memiliki N jenis kemungkinan konversi. Berdasarkan faktor kendala, faktor kesesuaian, dan kondisi lingkungan, kemungkinan konversi setiap sel harus diberikan nilai evolusi penggunaan lahan yang komprehensif. Jika nilai evaluasi arah transisi sel ditentukan oleh kemungkinan konversi terbesar, nilai tersebut dapat menyebabkan proliferasi tipe penggunaan lahan dominan dan menyebabkan simulasi tipe penggunaan lahan dominan yang berlebihan dan simulasi tipe penggunaan lahan lemah yang tidak memadai. Oleh karena itu, nilai evaluasi komprehensif penggunaan lahan dan matriks transfer area perubahan penggunaan lahan digunakan untuk menentukan arah konversi sel. Gambar 4.9 menunjukkan alur proses konversi tipe penggunaan lahan pada sel.

Langkah-langkahnya adalah sebagai berikut:

a. Menghitung nilai evaluasi maksimum dari tabel ringkasan statistik evaluasi komprehensif yang diperoleh dari Cloud-$CELUC$.
b. Loop membaca setiap baris tabel. Setiap baris merupakan pasangan nilai kunci $((i,j), CELUCs)$, dimana (i,j) adalah posisi sel, $CELUCs$ berarti sel (i,j) mempunyai

N jenis kemungkinan konversi penggunaan lahan $(CELUC)$, dan $CELUCi$ berarti $CELUC$ ke-i dari sel.

c. Menentukan apakah luasan tipe penggunaan lahan yang terkonversi mencapai batas atas luasan konversi tipe penggunaan lahan tersebut atau tidak.
   Batas atas konversi tipe penggunaan lahan diperoleh dari matriks probabilitas transisi area masing-masing tipe penggunaan lahan, didefinisikan sebagai pasangan nilai kunci (V_{mk}, P_{mk}) dimana V_{mk} adalah area tipe penggunaan lahan matriks konversi dan P_{mk} adalah matriks probabilitas transisi penggunaan lahan pada tahap CLOUD-Markov.
d. Jika mencapai batas area atas, $CELUCi$ sel harus ditandai dengan 0, artinya salah satu $CELUCi$ sel (i,j) telah dihapus untuk memastikan $CELUCi$ tidak digunakan pada langkah selanjutnya. Kemudian, ia kembali ke langkah pertama.



Gambar 4.9. Alur konversi tipe penggunaan lahan di sel.

Jika batas atas area tidak tercapai, tipe penggunaan lahan dari sel akan diubah menjadi tipe penggunaan lahan baru dan disimpan sebagai pasangan nilai kunci

$((i,j),CELUCi)$ ke dalam array, membuang $CELUC$ lain untuk memastikan sel tidak digunakan pada langkah selanjutnya. Kemudian, ia kembali ke langkah pertama.

e. Mengulangi langkah di atas hingga semua sel menyelesaikan konversi, dan akhirnya mendapatkan prediksi dari seluruh distribusi perubahan penggunaan lahan, yang disimpan sebagai sebuah array. Setiap item array adalah pasangan nilai kunci $((i,j),CELUC)$.

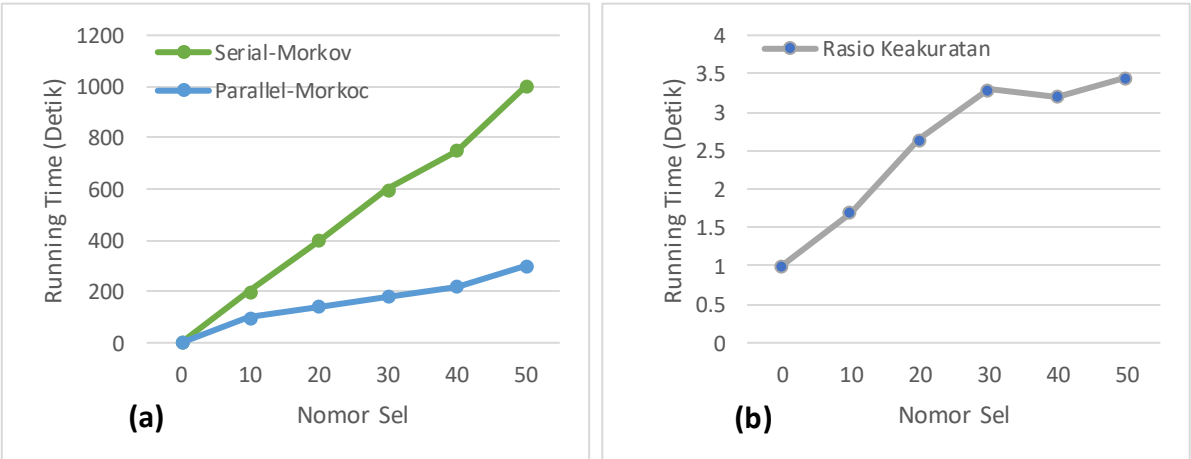
4.7 ANALISIS EFISIENSI MODEL

Satu mesin digunakan sebagai node master untuk pekerjaan NameNode dan JobTracker, dan empat mesin lainnya digunakan sebagai node budak untuk pekerjaan DataNode dan TaskTracker. Lingkungan operasi mesin adalah sistem CentOS 7.1.1503 dengan Java versi 1.8.0_112 dan Hadoop versi 2.7.3. Konfigurasi lingkungan percobaan ditunjukkan pada Tabel 4.4. Lingkungan perangkat keras algoritma serial sama dengan perangkat keras node Hadoop.

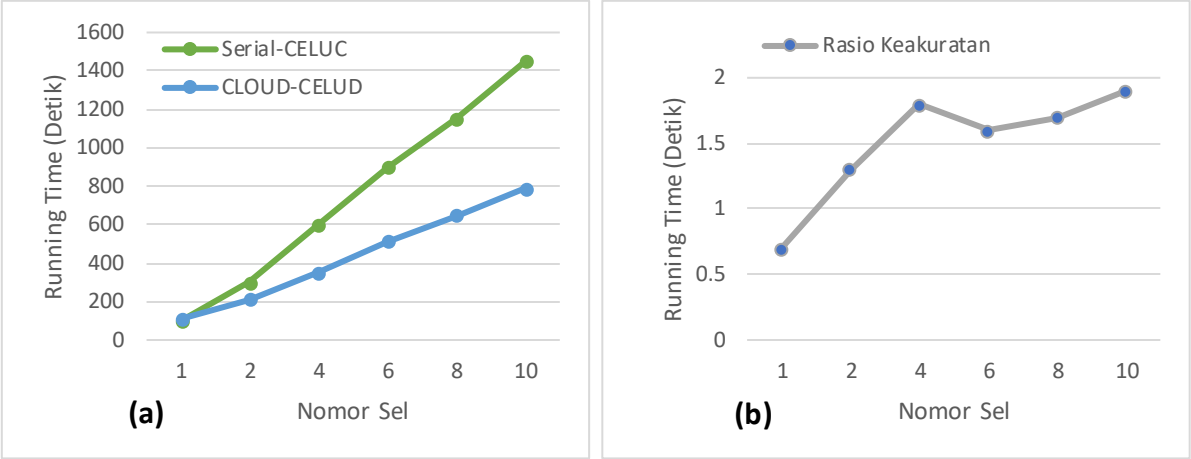
Tabel 4.4. Lingkungan percobaan Hadoop.

Alamat IP	Peran Node	CPU	RAM
192.168.128.1	Master/Namenode/JobTracker	Four-core 2,4 Ghz	4G
192.168.128.2	Slave/Datanode/TaskTracker	Four-core 2,4 Ghz	4G
192.168.128.3	Slave/Datanode/ TaskTracker	Four-core 2,4 Ghz	4G
192.168.128.4	Slave/Datanode/ TaskTracker	Four-core 2,4 Ghz	4G
192.168.128.5	Slave/Datanode/ TaskTracker	Four-core 2,4 Ghz	4G

Hasil rasio efisiensi dan percepatan running algoritma serial-Markov relatif terhadap algoritma Cloud-Markov ditunjukkan pada Gambar 4.10. Hasil rasio efisiensi dan akselerasi running algoritma serial-CELUC relatif terhadap algoritma Cloud-CELUC ditunjukkan pada Gambar 4.11. Seperti yang ditunjukkan pada Gambar 4.10a dan 4.11a, sumbu absis menunjukkan jumlah sel (1 n kira-kira 9.000.000) dan sumbu ordinat menunjukkan waktu berjalan.



Gambar 4.10. Rasio efisiensi dan akselerasi berjalan Cloud-Markov relatif terhadap serial-Markov. (a) Perbandingan efisiensi lari, (b) rasio percepatan.



Gambar 4.11. Rasio efisiensi dan akselerasi pengoperasian serial-CELUC relatif terhadap Cloud-CELUC. (a) Perbandingan efisiensi lari, (b) rasio percepatan.

Hasil penelitian menunjukkan bahwa waktu eksekusi Cloud-Markov lebih sedikit dibandingkan dengan algoritma serial Markov, dan dengan bertambahnya data masukan maka rasio percepatannya meningkat dan cenderung lancar. Rasio akselerasi algoritma Cloud-Markov terhadap algoritma serial Markov cenderung stabil pada angka 3,27, dan rasio akselerasi Cloud-CELUC terhadap serial CELUC cenderung stabil pada angka 1,77. Rasio akselerasi tertinggi Cloud-Markov bisa mencapai 3,43, dan rasio akselerasi tertinggi Cloud-CELUC bisa mencapai 1,86.

Efisiensi model Markov paralel berdasarkan lingkungan cloud sangat luar biasa karena sistem MapReduce secara efektif mendistribusikan beban kerja dua fase pencocokan sel dan statistik kuantitatif. Efisiensi Cloud-CELUC juga ditingkatkan karena fase Peta yang kami definisikan berjalan sangat cepat. Akan tetapi, ketika keluaran dari fase Peta dimasukkan ke dalam fase Pengurangan, hal ini menghabiskan sebagian besar waktu berjalan yang panjang, sehingga mengurangi efisiensi operasi.

Evaluasi Presisi dan Analisis Hasil

Citra penginderaan jauh tahun 2006 dan data lainnya ditetapkan sebagai data awal, dan perubahan penggunaan lahan tahun 2013 disimulasikan dengan menggugat model paralel CA-Markov. Dalam percobaan kami, luas perairan ditetapkan tetap dan tidak ada perubahan. Berdasarkan data penggunaan lahan tahun 2006 dan 2013, matriks probabilitas transisi kawasan dapat dihitung. Tabel 4.5 adalah matriks transisi wilayah tahun 2006–2013, yang setiap selnya mewakili total luas suatu tipe penggunaan lahan yang berpindah ke tipe penggunaan lahan lain dari tahun 2006 hingga 2016. Tabel 4.6 adalah matriks probabilitas transisi tahun 2006–2013, di yang setiap sel mewakili kemungkinan perpindahan suatu tipe penggunaan lahan ke tipe penggunaan lahan lainnya dari tahun 2006 hingga 2016.

Tabel 4.5. Matriks transisi wilayah tahun 2006–2013 (satuan: km2).

2006 \ 2013	Lahan pertanian	Lahan kontruksi	Cagar alam	Total
Lahan pertanian	1282.95	409.71	95.67	1788.33

Tanah kontruksi	210.94	1381.69	12.52	1605.15
Lahan cagar alam	93.39	50.79	4409.60	4553.78
Total	1587.28	1842.19	4517.79	7947.26

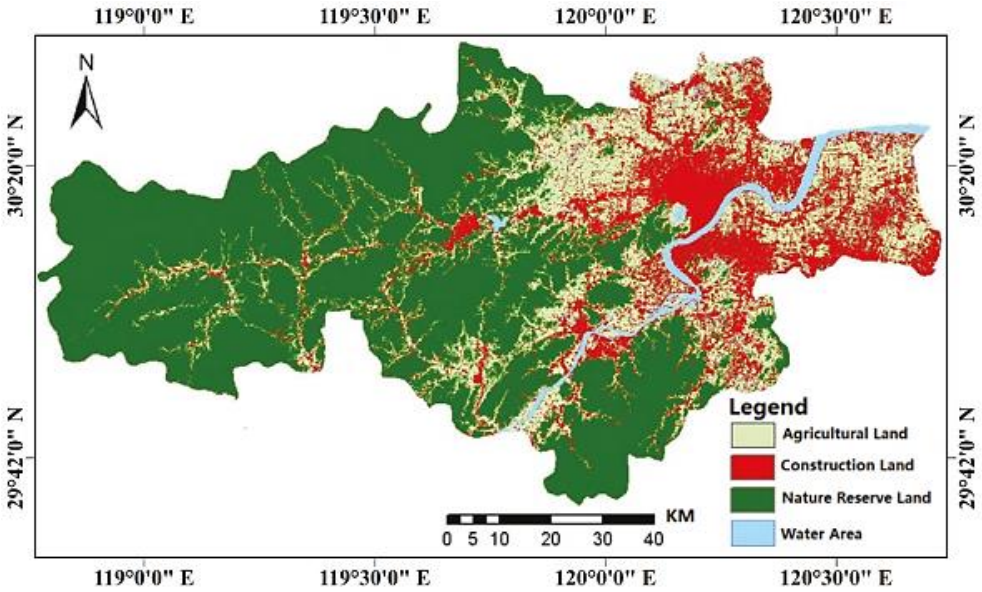
Tabel 4.6. Matriks probabilitas transisi tahun 2006–2013 (satuan:%).

2006 \ 2013	Lahan pertanian	Lahan kontruksi	Cagar alam
Lahan pertanian	71.74	22.91	5.35
Tanah kontruksi	13.14	86.08	0.78
Lahan cagar alam	2.05	1.12	6.839

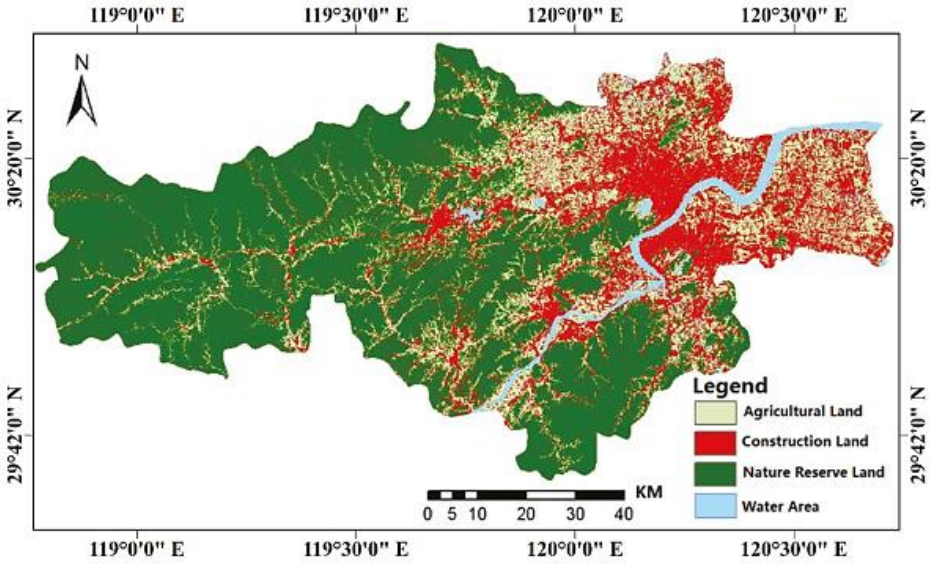
Seperti ditunjukkan pada Tabel 4.5 dan 4.6, transisi tipe penggunaan lahan terbesar adalah peralihan lahan pertanian ke lahan konstruksi; rasionya mencapai 22,91%. Untuk mengevaluasi presisi yang disimulasikan, data penggunaan lahan tahun 2013 diklasifikasikan menggunakan citra penginderaan jauh tahun 2013 yang sebenarnya pada tahap pemrosesan data. Data simulasi penggunaan lahan dan data klasifikasi penggunaan lahan tahun 2013 masing-masing ditunjukkan pada Gambar 4.12 dan 4.13.

Setelah simulasi, percobaan evaluasi presisi dilakukan untuk mengoreksi semua jenis parameter bobot yang ditentukan dalam C-MCE. Setelah sejumlah besar percobaan berulang dan koreksi parameter bobot, parameter bobot faktor kesesuaian diperoleh, yang tercantum dalam Tabel 4.2. Saat ini, metode evaluasi presisi yang umum digunakan meliputi perbandingan visual, uji dimensi, kontras piksel, dan Uji koefisien Kappa.

Dengan perbandingan visual, diperoleh simulasi cagar alam di wilayah barat dan selatan yang paling akurat. Dapat disimpulkan bahwa faktor kesesuaian dan faktor penghambat sejalan dengan tren perubahan cagar alam di wilayah penelitian. Lahan konstruksi di pusat kota memiliki akurasi simulasi yang lebih baik. Namun lahan pembangunan di timur dan utara, termasuk kota Yuanpu dan Linpu, tersebar dan terjalin dengan lahan pertanian. Oleh karena itu, kesalahan simulasi relatif besar.



Gambar 4.12. Peta simulasi penggunaan lahan tahun 2013



Gambar 4.13. Peta klasifikasi citra penginderaan jauh tahun 2013.

Uji koefisien Kappa adalah metode uji kuantitatif yang paling umum digunakan. Ketika koefisien Kappa digunakan untuk membandingkan konsistensi data, kriteria yang umum digunakan adalah sebagai berikut: Jika kedua peta penggunaan lahan tersebut identik, maka Kappa = 1; ketika Kappa > 0,8, konsistensinya hampir sempurna; ketika 0,6 < Kappa ≤ 0,8, konsistensi cukup besar; ketika 0,4 < Kappa ≤ 0,6, konsistensinya sedang; ketika 0,2 < Kappa ≤ 0,4, konsistensinya sedikit; ketika 0 < Kappa ≤ 0,2, konsistensinya buruk.

Data simulasi penggunaan lahan dan data penggunaan lahan aktual tahun 2013 dibandingkan, dan hasil koefisien Kappa masing-masing ditunjukkan pada Tabel 4.7–4.9.

Tabel 4.7. Tabel uji koefisien Kappa lahan cagar alam tahun 2013 (satuan: km2).

Data Simulasi / Data Rahasia	Cagar Alam	Lahan Non-Cagar alam	Total	Ketepatan	Kappa
Lahan Cagar Alam	4221.35	288.47	4509.82	93.60%	
Lahan Non-Cagar Alam	296.44	3421.50	3727.94	92.05%	0,86
Total	4517.79	3717.97			

Tabel 4.8. Koefisien Kappa lahan pembangunan tahun 2013 (satuan: km2).

Data Simulasi / Data Rahasia	Lahan Kontruksi	Lahan Non-Kontruksi	Total	Ketepatan	Kappa
Lahan Kontruksi	1391.41	452.77	1844.18	75.45%	
Lahan Non-Kontruksi	450.78	5942.80	6393.58	92.95%	0,86
Total	1842.19	6395.57			

Tabel 4.9. Koefisien Kappa lahan pertanian tahun 2013 (satuan: km2).

Data Simulasi / Data Rahasia	Lahan Pertanian	Lahan Non-Pertanian	Total	Ketepatan	Kappa
Lahan pertanian	1152.64	427,17	1579.81	72.96%	
Lahan Non-Pertanian	434.65	6223.30	6657.95	93.47%	0,66
Total	1587.29	6650.47			

Hasil penelitian menunjukkan bahwa koefisien Kappa untuk cagar alam, lahan konstruksi, dan lahan pertanian masing-masing adalah 0,85, 0,6, dan 0,65. Ini berarti bahwa hasil simulasi tahun 2013 cukup akurat, dan penggunaan model paralel CA-Markov untuk memprediksi penggunaan lahan di masa depan akan sangat andal.

4.8 PREDIKSI PERUBAHAN PENGGUNAAN LAHAN

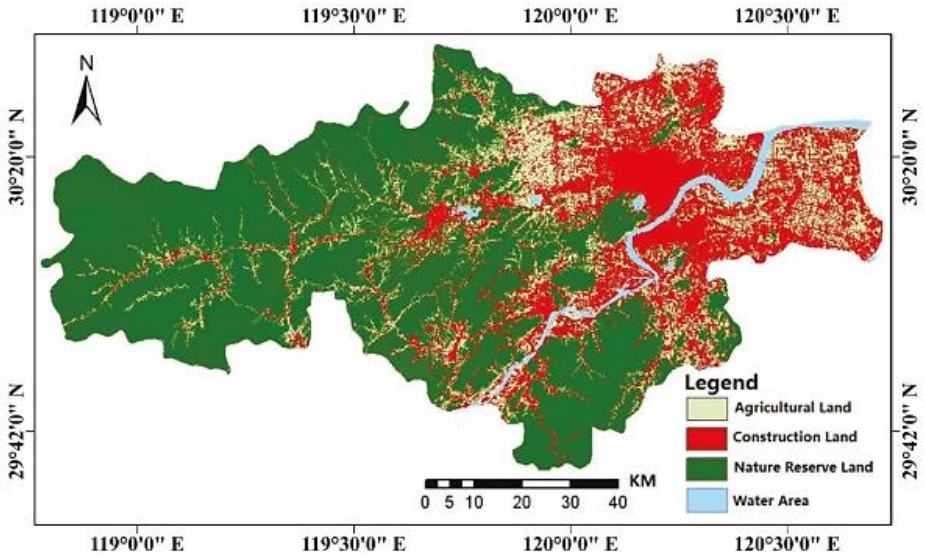
Berdasarkan data klasifikasi penggunaan lahan tahun 2013 dan data eksperimen lainnya, perubahan penggunaan lahan tahun 2020 diprediksi menggunakan model paralel CA-Markov. Hasil matriks transisi kawasan tahun 2013–2020 disajikan pada Tabel 4.10, dan peta prediksi penggunaan lahan tahun 2020 disajikan pada Gambar 4.14.

Tabel 4.10. Matriks transisi kawasan tahun 2013–2020 (satuan: km2).

2013 \ 2020	Lahan pertanian	Lahan kontruksi	Cagar alam	Total
Lahan pertanian	1133.35	361.94	84.52	1579.81
Tanah kontruksi	242.35	1587.45	14.38	1844.18
Lahan cagar alam	92.49	50.30	4367.03	4509.82
Total	1468.19	1999.69	4465.93	7933.81

Terlihat dari peta prediksi penggunaan lahan tahun 2020, lahan konstruksi di wilayah studi secara keseluruhan mengalami peningkatan, dan peningkatan ini terutama disebabkan oleh konversi lahan pertanian. Lahan pertanian menunjukkan tren penurunan, dan lahan cagar alam tidak banyak mengalami perubahan dibandingkan dengan luasnya.

Pertumbuhan lahan konstruksi secara keseluruhan relatif besar. Secara khusus, karena perluasan sistem jalan raya dan transportasi umum, lahan konstruksi tumbuh lebih cepat di pusat kota di setiap daerah. Pertumbuhan lahan konstruksi di distrik Xihu, Gongshu, Xiacheng, Binjiang, dan Xiaoshan sangat menonjol. Karena keterbatasan medan dan badan air, kabupaten dan wilayah perkotaan lainnya mempertahankan areal lahan konstruksi yang stabil.



Gambar 4.14. Peta prediksi penggunaan lahan tahun 2020.

Lahan pertanian dan lahan konstruksi saling terkait di wilayah yang luas di barat laut dan utara wilayah penelitian. Namun nilai evaluasi lahan konstruksi di kawasan ini tidak tinggi karena jauh dari pusat kota dan pusat kabupaten serta sistem transportasi umum yang kurang berkembang. Luas lahan pertanian di wilayah tersebut pada dasarnya stabil. Lahan pertanian di sisa wilayah studi dipengaruhi oleh perluasan perkotaan, jalan raya, dan sistem transportasi umum, dan sebagian besar lahan diubah menjadi lahan konstruksi.

Cagar alam sebagian besar terdapat di kawasan perbukitan bagian barat dan selatan, yang mengalami berbagai kendala dalam pembangunan, seperti medan yang landai dan pembatasan perlindungan ekologi, serta sistem transportasi yang tidak nyaman, sehingga pada dasarnya arealnya tetap stabil.

4.9 RINGKASAN

Eksperimen menunjukkan bahwa hasil simulasi penggunaan lahan berdasarkan model CA-Markov di lingkungan awan dapat diandalkan, yang mencerminkan bahwa metode yang diusulkan dalam penelitian ini dapat diandalkan dan dapat diterapkan. Sementara itu, MapReduce efektif dalam memparalelkan model CA-Markov untuk meningkatkan kecepatan pemrosesan prediksi perubahan penggunaan lahan berdasarkan model CA-Markov. Metode ini memparalelkan model CA-Markov menjadi dua bagian: Model Markov paralel berdasarkan lingkungan cloud (Cloud-Markov), dan metode evaluasi komprehensif perubahan penggunaan lahan berdasarkan MapReduce (Cloud-CELUC). Dengan memilih Hangzhou sebagai wilayah studi dan menyiapkan lingkungan eksperimen Hadoop, eksperimen tersebut dirancang untuk memverifikasi keandalan, presisi, dan efisiensi pengoperasian metode tersebut. Perubahan penggunaan lahan di Hangzhou pada tahun 2020 disimulasikan dan hasilnya dianalisis. Hasil percobaan menunjukkan bahwa metode yang sekaligus mewujudkan integritas dan segmentasi untuk simulasi dan prediksi perubahan penggunaan lahan juga praktis dan efektif.

Penelitian ini telah berhasil menerapkan kerangka MapReduce untuk meningkatkan efisiensi prediksi perubahan penggunaan lahan. Namun, masih ada beberapa masalah penting yang perlu diselidiki lebih lanjut. Pertama, perubahan penggunaan lahan tidak hanya dibatasi oleh kondisi alam, namun juga oleh faktor politik, ekonomi, demografi, dan faktor kompleks lainnya. Karena keterbatasan sumber data, penelitian ini membangun modelnya terutama berdasarkan faktor lalu lintas, medan, dan lokasi. Jika lebih banyak data tersedia, modul prediksi pola spasial penggunaan lahan, sosial, ekonomi, kebijakan, demografi, dan faktor lainnya harus diperhitungkan dalam penelitian di masa depan. Kami juga akan mempertimbangkan untuk menggabungkan metode identifikasi gambar otomatis ke dalam pekerjaan kami saat ini untuk mengurangi pekerjaan prapemrosesan manual. Ketika Cloud-CELUC memperoleh hasil output dari tahap Map di tahap Reduce, input/output (IO) menjadi penghambat kinerja sistem. Oleh karena itu, upaya penelitian lebih lanjut harus didedikasikan untuk menguji peningkatan efisiensi IO pada kinerja komputasi awan berdasarkan model CA-Markov.

BAB 5
ANALISIS MEDAN DI MESIN GOOGLE EARTH

Analisis medan merupakan alat penting untuk memodelkan sistem lingkungan. Bertujuan untuk menggunakan kemampuan komputasi berbasis awan dari Google Earth Engine (GEE), kami menyesuaikan algoritme untuk menghitung atribut medan, seperti kemiringan, aspek, dan kelengkungan, untuk resolusi dan cakupan geografis yang berbeda. Metode penghitungannya didasarkan pada nilai geometri dan ketinggian yang diperkirakan dalam jendela bola 3 × 3, dan tidak bergantung pada data ketinggian yang diproyeksikan. Dengan demikian, turunan parsial dari medan dihitung dengan mempertimbangkan jarak lingkaran besar dari titik referensi permukaan topografi. Algoritme ini dikembangkan menggunakan antarmuka pemrograman JavaScript dari editor kode online GEE dan dapat dimuat sebagai paket khusus. Algoritme ini juga menyediakan fitur tambahan untuk membuat visualisasi peta medan dengan skala legenda dinamis, yang berguna untuk memetakan wilayah yang berbeda: dari lokal ke global. Kami membandingkan konsistensi metode yang diusulkan dengan alat analisis medan GEE yang tersedia namun terbatas, yang menghasilkan korelasi masing-masing sebesar 0,89 dan 0,96 untuk aspek dan kemiringan dalam skala hampir global. Selain itu, kami membandingkan kemiringan, aspek, kelengkungan horizontal, dan vertikal suatu situs referensi (Gunung Ararat) dengan atribut setara yang diperkirakan pada Sistem Analisis Geografis Otomatis (SAGA), yang menghasilkan korelasi antara 0,96 dan 0,98. Korespondensi visual TAGEE dan SAGA menegaskan potensinya untuk analisis medan. Algoritme yang diusulkan dapat berguna untuk membuat analisis medan terukur dan disesuaikan dengan kebutuhan yang disesuaikan, memanfaatkan antarmuka GEE yang berkinerja tinggi.

5.1 PENDAHULUAN

Analisis medan sangat penting untuk memodelkan sistem lingkungan. Variabilitas bentang alam sering digunakan untuk memahami, memetakan atau memodelkan proses geomorfologi, hidrologi, dan biologis. Ketinggian mempunyai hubungan yang kuat dengan suhu terestrial, jenis vegetasi, dan energi potensial yang terakumulasi pada suatu lereng. Aspek dan produk turunannya, seperti atribut Keutaran dan Ketimuran, dapat dikaitkan dengan potensi penyinaran matahari di suatu daerah. Gradien lereng, misalnya, mengontrol kecepatan aliran permukaan dan bawah permukaan serta laju limpasan. Demikian pula, kelengkungan berhubungan dengan percepatan dan dispersi aliran air dan sedimen, yang berdampak pada erosi dan kandungan air tanah.

Ketersediaan publik atas data ketinggian dengan cakupan global, seperti model ketinggian digital (DEM) yang berasal dari Shuttle Radar Topography Mission (SRTM DEM) milik NASA dan model permukaan digital dari Advanced Land Observing Satellite (AW3D30 DSM), telah mendorong eksplorasi fitur topografi dalam konteks berbeda menggunakan alat pemrosesan yang tersedia di beberapa sistem informasi geografis (GIS). Namun, meskipun

banyak kumpulan data ketinggian global telah dipopulerkan, penting untuk memperhatikan kualitasnya ketika digunakan untuk tujuan pemodelan, karena rata-rata perolehan dan aspek produksi lainnya dapat berdampak signifikan terhadap keluaran. Selain itu, analisis kumpulan data geografis yang besar dapat menimbulkan beberapa keterbatasan pada GIS tradisional. Hal ini menjadi lebih penting dengan tersedianya kumpulan data digital baru, yang memberikan resolusi temporal dan spasial yang lebih baik karena kemajuan teknologi sensor.

Data Ketinggian Medan Multi-resolusi Global 2010 dan setelan atribut medan global adalah contoh kumpulan data yang dihasilkan menggunakan tugas komputasi besar untuk memetakan cakupan global dan dalam resolusi spasial yang berbeda, yang memerlukan arsitektur pemrosesan yang dioptimalkan. Secara umum, arsitektur berkinerja tinggi didasarkan pada pemisahan data menjadi subset (ubin) yang lebih kecil untuk memanfaatkan operasi komputasi terdistribusi. Baru-baru ini, dengan munculnya dan mempopulerkan antarmuka berbasis awan untuk memproses data geografis berukuran besar, misalnya Google Earth Engine, paket perangkat lunak Pangeo, dan layanan Actinia REST, tugas komputasi yang diterapkan pada analisis medan dapat dapat diskalakan dan disesuaikan secara langsung oleh pengguna.

Earth Engine (GEE) adalah platform berbasis cloud yang dikembangkan oleh Google yang mendukung analisis katalog besar data Observasi Bumi dalam skala global. Data ini telah digunakan untuk memetakan perubahan hutan global pada abad ke-21, perubahan air permukaan bumi, wilayah perkotaan global, perkembangan kebakaran hutan, perubahan permukaan bumi yang gundul, dan lain-lain. Dalam hal ini, GEE menjadi menarik bukan karena tugas pemrosesan terdistribusi dijalankan di sisi server Google, namun juga karena meningkatnya ketersediaan banyak kumpulan data geografis global yang dapat dieksplorasi dalam pemetaan topografi. Terdapat beberapa data topografi yang tersedia dalam GEE, seperti SRTM DEM global, AW3D30 DSM, data Global 30 Arc-Second Elevation (GTOPO30 DEM), dan lain-lain. Dengan demikian, karakteristik GEE dapat memungkinkan penyesuaian analisis medan berkinerja tinggi dengan input pengguna minimal dan pemrosesan komputasi apa pun di sisi pengguna. Faktanya, GEE menyediakan tiga algoritma untuk menghitung kemiringan, iluminasi, dan aspek medan, namun kurang menyediakan metode penghitungan informasi medan lainnya, seperti kelengkungan dan karakterisasi lanskap.

Selain itu, kendala umum dalam analisis medan global dalam GIS umum adalah kebutuhan untuk memproyeksikan DEM ke dalam sistem koordinat yang diproyeksikan, yang memastikan data ketinggian memiliki jarak yang sama pada kotak bidang persegi. Langkah ini rumit karena sulit untuk menentukan sistem proyeksi yang meminimalkan distorsi medan secara global. Selain itu, karena banyak DEM global yang tersedia direferensikan oleh sistem koordinat geografis dan beberapa peneliti terus menerapkan algoritma grid persegi pada sistem tersebut, algoritma tersebut harus mempertimbangkan geometri dan spesifisitas DEM spheroidal global. Aspek ini penting karena penerapan metode grid persegi pada DEM sudut sama sferoidal menyebabkan kesalahan komputasi yang besar dalam model variabel morfometrik.

Dalam buku ini, penulis bertujuan untuk mendeskripsikan dan menyediakan algoritma pemrosesan yang mudah digunakan untuk melakukan analisis medan di GEE. Algoritme ini memanfaatkan arsitektur berkinerja tinggi GEE untuk membuat analisis komputasi dapat diskalakan, disesuaikan dengan kebutuhan yang disesuaikan, dan memerlukan masukan pengguna yang minimal. Untuk hal ini, paket yang diusulkan memanfaatkan metode perhitungan yang diadaptasi untuk grid elevasi spheroidal, yang mendukung analisis skala global terhadap resolusi DEM yang berbeda tanpa memproyeksikan data elevasi.

5.2 DESKRIPSI ALGORITMA TERRAIN ANALYSIS IN GEE (TAGEE)

Paket Terrain Analysis in GEE (TAGEE) menggunakan metode perhitungan yang disesuaikan dengan grid sudut spheroidal, yaitu DEM dapat direferensikan dalam sistem koordinat geografis, misalnya World Geodetic System (WGS84). Paragraf berikut menjelaskan secara singkat metode penghitungan yang dilakukan oleh paket TAGEE. Pembaca dirujuk untuk konsep matematika geomorfometri, gambaran sejarah kemajuan pemodelan medan digital, dan gagasan permukaan topografi dan keterbatasannya.

Permukaan Topografi

Topografi daratan dapat didekati dengan permukaan topografi yang ditentukan oleh fungsi bivariat bernilai tunggal yang kontinu (Persamaan (1)):

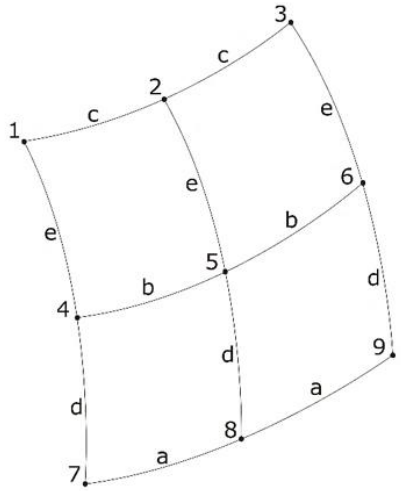
$$z = f(x,y)$$

dimana z adalah ketinggian (meter), dan x dan y adalah koordinat dalam koordinat geografis (derajat). Variabel morfometri lokal merupakan fungsi dari turunan parsial ketinggian. Menggunakan Metode Evans – Young, fungsi $z = f(x,y)$ dinyatakan sebagai polinomial Taylor bivariat orde kedua (Persamaan (2)):

$$z = \frac{rx^2}{2} + \frac{ty^2}{2} + sxy + px + qy + u$$

dimana r, t, s, p, dan q adalah turunan parsial, dan u adalah suku sisa.

Berbeda dari model elevasi digital yang diproyeksikan pada kotak bidang persegi, yang mana turunan parsial medan diperkirakan dengan perbedaan berhingga, pemrosesan dan analisis DEM sudut sama sferoidal harus mempertimbangkan geometri sferoidal. Dalam kasus seperti ini, jarak grid dengan unit linier yang kira-kira sama sepanjang meridian dan paralel hanya ada di ekuator. Untuk memperkirakan parameter grid spheroidal, jendela bergerak 3 × 3 harus mengambil elemen geometri dan nilai elevasi dari node jendela (Gambar 5.1).



Gambar 5.1. Kotak bersudut sama besar berbentuk bola berukuran 3 × 3 dengan geometri linier $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$, dan $\mathbf{f}$, dan sembilan titik ketinggian.

Parameter Medan: Ketinggian dan Geometri Tetangga

Nilai ketinggian jendela bergerak 3 × 3 diperkirakan dengan kernel konvolusi. Untuk geometri, rumus Haversine digunakan untuk menentukan jarak lingkaran besar antara dua titik tetangga dalam jendela bola, berdasarkan posisi geografis lintang dan bujurnya (Persamaan (3)–(5)):

$$j = sin^2\left(\frac{\Delta\emptyset}{2}\right) + cos\ \emptyset_1 \cdot \cos\emptyset_2 \cdot sin^2\frac{\Delta\lambda}{2}$$

$$k = 2 \cdot atan2\left(\sqrt{j}, \sqrt{(1-j)}\right)$$

$$l = R \cdot k$$

dimana $\emptyset_1$ adalah garis lintang untuk titik tertentu yang pertama dalam radian, $\emptyset_1$ adalah garis lintang untuk titik tertentu yang kedua dalam radian, λ_1 adalah garis bujur untuk titik tertentu yang pertama dalam radian, λ_2 adalah garis bujur untuk titik tertentu yang kedua dalam radian, $\Delta\phi$ dan $\Delta\lambda$ adalah perbedaan lintang dan bujur antara titik-titik tertentu, dan R adalah radius rata-rata Bumi yang sama dengan 6.371.000 meter. Jarak linier l diberikan dalam meter.

Mengetahui garis lintang dan bujur dari simpul jendela (Gambar 5.1), rumus Haversine memungkinkan penghitungan jarak linier $a, b, c, d,$ dan $e,$ yang digunakan dengan nilai ketinggian tetangga (dari z_1 hingga z_9) ke menghitung turunan parsial medan.

Turunan Medan

Untuk memperkirakan turunan parsial orde pertama dan kedua r, t, s, p dan q, model polinomial dilengkapi dengan kuadrat terkecil dan menghasilkan estimasi berikut (Persamaan (6)–(10)):

$$p = \frac{a^2cd(d+e)(z_3 - z_1) + b(a^2d^2 + c^2e^2)(z_6 - z_4) + ac^2e(d+e)(z_9 - z_7)}{2[a^2c^2(d+e)^2 + b^2(a^2d^2 + c^2e^2)]}$$

$$q = \frac{1}{3de(d+e)(a^4+b^4+c^4)} \cdot \begin{Bmatrix} [d^2(a^4+b^4+b^2c^2)+c^2e^2(a^2-b^2)](z_1+z_3) \\ -[d^2(a^4+c^4+b^2c^2)-e^2(a^4+c^4+a^2b^2)](z_4+z_6) \\ -[e^2(b^4+c^4+a^2b^2)-a^2d^2(b^2-c^2)](z_7+z_9) \\ +d^2[b^4(z_2-3z_5)+c^4(3z_2-z_5)+(a^4-2b^2c^2)(z_2-z_5)] \\ +e^2[a^4(z_5-3z_8)+b^4(3z_5-z_8)+(c^4-2a^2b^2)(z_5-z_8)] \\ -2[a^2d^2(b^2-c^2)z_8+c^2e^2(a^2-b^2)z_2] \end{Bmatrix}$$

$$r = \frac{c^2(z_1+z_3-2z_2)+b^2(z_4+z_6-2z_5)+a^2(z_7+z_9-2z_8)}{a^4+b^4+c^4}$$

$$s = \{c(a^2[(d+e)+b^2e])(z_3-z_1)-b(a^2d-c^2e)(z_4-z_6)+a[c^2(d+e)+b^2d](z_7-z_9)\} \cdot \frac{1}{2[a^2c^2(d+e)^2+b^2(a^2d^2+c^2e^2)]}$$

$$t = \frac{2}{3ed(d+e)(a^4+b^4+c^4)} \cdot \begin{Bmatrix} [d(a^4+b^4+b^2c^2)-c^2e(a^2-b^2)](z_1+z_3) \\ -[d(a^4+c^4+b^2c^2)+e(a^4+c^4+a^2b^2)](z_4+z_6) \\ +[e(b^4+c^4+a^2b^2)+a^2d(b^2-c^2)](z_7+z_9) \\ +d[b^4(z_2-3z_5)+c^4(3z_2-z_5)+(a^4-2b^2c^2)](z_2-z_5) \\ +e[a^4(3z_8-z_5)+b^4(z_8-3z_5)+(c^4-2a^2b^2)](z_8-z_5) \\ -2[a^2d(b^2-c^2)z_8-c^2e(a^2-b^2)z_2] \end{Bmatrix}$$

dimana parameter a, b, c, d, dan e adalah jarak linier yang dihitung dari rumus Haversine (Persamaan (3)–(5)), dan nilai z adalah nilai elevasi dari tetangga jendela bergerak (Gambar 5.1).

Atribut Medan

Atribut lokal, seperti kemiringan, aspek, dan kelengkungan, dihitung dari turunan parsial medan. Gradien kemiringan (G, Persamaan (11)) adalah atribut aliran yang berhubungan dengan kecepatan aliran yang digerakkan oleh gravitasi. Untuk mengukur arah digunakan aspek kemiringan (A, Persamaan (12) dan (13)). Selain itu, seseorang dapat menghitung besarnya kemiringan yang dihadapi ke arah Utara atau Timur, sehingga menghasilkan Keutaran (A_N, Persamaan (14)) dan Ketimuran (A_E, Persamaan (15)) yang diperoleh dari aspek tersebut. Atribut fluks sisa yang dapat dihitung dari turunan parsial orde pertama dan kedua adalah kelengkungan horizontal (k_h, Persamaan (16)) dan kelengkungan vertikal (k_v, Persamaan (17)). Meskipun kelengkungan horizontal menunjukkan apakah aliran lateral menyatu ($k_h > 0$) atau menyimpang ($k_h > 0$), kelengkungan vertikal mengukur percepatan relatif ($k_v > 0$) dan perlambatan ($k_v > 0$) dari aliran yang digerakkan oleh gravitasi:

Persamaan 11 $$G = \arctan\sqrt{p^2 + q^2}$$

Persamaan 12 $$-90[1 - \text{sign}(q)](1 - |\text{sign}(p)|) + 180[1 + \text{sign}(p)] - \frac{180}{\pi}\text{sign}(p)\arccos\left(\frac{-q}{\sqrt{p^2 + q^2}}\right)$$

Persamaan 13 $$\text{sign}(x)\begin{cases} 1 & \text{for} x > 0 \\ 0 & \text{for} x = 0 \\ -1 & \text{for} x < 0 \end{cases}$$

Persamaan 14 $$A_N = \cos A$$

Persamaan 15 $$A_E = \sin A$$

Persamaan 16 $$k_h = \frac{q^2 r + 2pqs + p^2 t}{(p^2 + q^2)\sqrt{1 + p^2 + q^2}}$$

Persamaan 17 $$k_v = \frac{p^2 r - 2pqs + q^2 t}{(p^2 + q^2)\sqrt{(1 + p^2 + q^2)^3}}$$

Berbeda dari atribut aliran, yang merupakan variabel spesifik medan gravitasi, atribut bentuk berkaitan dengan bagian utama medan. Kelengkungan rata-rata (H, Persamaan (18)) adalah setengah jumlah dari dua bagian normal ortogonal dan mewakili dua mekanisme akumulasi aliran yang digerakkan oleh gravitasi dengan bobot yang sama: konvergensi dan perlambatan relatif. Di antara kelas atribut bentuk, kelengkungan Gaussian (K, Persamaan (19)) merupakan hasil kali kelengkungan maksimal (k_{max}) dan minimal (k_{min}). Dua kelengkungan utama menghitung kelengkungan tertinggi dan terendah untuk suatu titik tertentu pada permukaan topografi. Kelengkungan maksimal (k_{max}, Persamaan (20)) berguna untuk memetakan rigde ($k_{max} > 0$) dan depresi tertutup ($k_{max} < 0$). Demikian pula, kelengkungan minimal (k_{min}), Persamaan (21)) berguna untuk mengidentifikasi bukit ($k_{min} > 0$)dan lembah ($k_{min} < 0$) pada permukaan topografi. Dengan hasil kelengkungan mean dan Gaussian, klasifikasi bentuk lahan dapat dihasilkan setelah mengusulkan bentuk klasifikasi Gaussian yang berkelanjutan. Alih-alih memberikan nilai kategorikal, indeks bentuk (SI, Persamaan (22)) berkisar antara –1 hingga 1 dan peta bentuk lahan cembung ($SI > 0$) dan cekung ($SI < 0$)

Persamaan 18: $$H = -\frac{(1 + q^2)r - 2pqs + (1 + p^2)t}{2\sqrt{(1 + p^2 + q^2)^3}}$$

Persamaan 19: $$K = \frac{rt - S^2}{(1 + p^2 + q^2)^2}$$

Persamaan 20: $$k_{max} = H + \sqrt{(H^2 - K)}$$

Persamaan 21: $k_{min} = H - \sqrt{(H^2 - K)}$

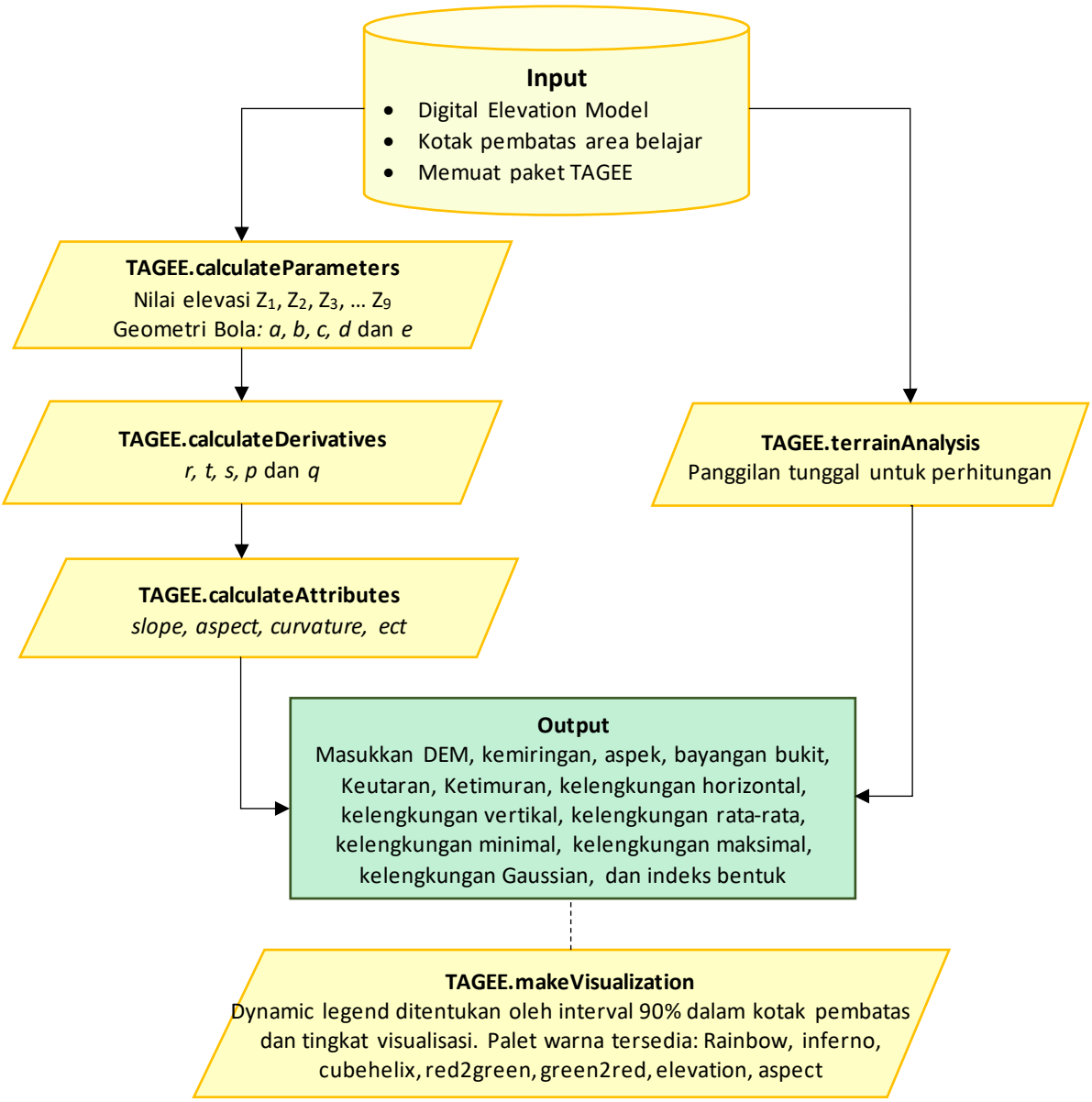
Persamaan 22: $SI = \frac{2}{\pi} \arctan \frac{H}{\sqrt{H^2 - K}}$

5.3 DESKRIPSI PAKET

Metode perhitungan yang disajikan dalam buku ini dikembangkan menggunakan antarmuka pemrograman JavaScript yang tersedia sebagai editor kode online GEE. TAGEE dikembangkan dengan modul perhitungan yang berbeda, serupa dengan apa yang dijelaskan dalam Metode. Modul pertama, calculParameters, menggunakan kernel konvolusi dan rumus Haversine untuk mengambil nilai ketinggian dan geometri bola dari jendela bergerak 3 × 3. Dalam modul ini, model elevasi digital dan poligon persegi yang mewakili kotak pembatas (min. Bujur, min. Lintang, maks. Bujur, dan maks. Lintang, dalam sistem referensi koordinat WGS84) diperlukan sebagai parameter masukan untuk dijalankan. Kotak pembatas digunakan dalam modul ini dan modul lainnya untuk menghasilkan gambar dengan nilai konstan dan membatasi perhitungan pada area studi. Modul pertama mengembalikan gambar dengan 14 band, yaitu nilai elevasi tetangga (dari z_1 hingga z_9) dan jarak (a, b, c, d, dan e) (Gambar 1).

Setelah parameter dasar (ketinggian dan jarak) ditetapkan, turunan parsial medan dihitung dengan modul calculDerivatives. Modul kedua ini memerlukan parameter yang dikembalikan dari calculParameters dan juga kotak pembatas wilayah studi. Modul kedua menambahkan turunan parsial (r, t, s, p, dan q) sebagai pita baru pada gambar sebelumnya. Kemudian, atribut medan dihitung dengan modul countAttributes (Gambar 5.2).

Atribut medan juga dapat dihitung dengan satu fungsi, tanpa memanggil modul perantara. Output akhir, untuk kedua alternatif (Gambar 5.2), adalah objek multi band yang berisi properti data yang sama dengan model elevasi digital (resolusi, tipe data, dan sistem referensi koordinat) dengan 13 band (Tabel 5.1). Atribut akhir dapat digunakan untuk pemodelan lebih lanjut di dalam GEE atau pemetaan tematik.



Gambar 5.2. Modul TAGEE untuk menghitung parameter medan, turunan, dan atribut.

Tabel 5.1. Atribut medan beserta satuan dan deskripsinya, dihitung dengan paket TAGEE.

Atribut	Satuan	Keterangan
Ketinggian	meter	Ketinggian medan di atas permukaan laut
Lereng	derajat	Gradien lereng
Aspek	derajat	Arah kompas
bayangan bukit	tak berdimensi	Kecerahan area yang diterangi
Keutuhan	tak berdimensi	Derajat orientasi ke Utara
ketimuran	tak berdimensi	Derajat orientasi ke Timur
Kelengkungan horisontal	meter	Lengkungan bersinggungan dengan garis kontur
Kelengkungan vertikal	meter	Lengkungan bersinggungan dengan garis lereng
Berarti kelengkungan	meter	Setengah jumlah dua kelengkungan ortogonal
Kelengkungan minimal	meter	Nilai kelengkungan terendah

Kelengkungan maksimal	meter	Nilai kelengkungan tertinggi
Kelengkungan Gaussian	meter	Hasil kali kelengkungan maksimal dan minimal
Bentuk Indeks	tak berdimensi	Bentuk klasifikasi Gaussian yang berkelanjutan

Paket tersebut memiliki fitur tambahan yang memudahkan visualisasi atribut medan. Karena kisaran nilai atribut dan resolusi piksel dapat bervariasi sesuai dengan tingkat visualisasi (zoom), yang berdampak pada perkiraan geometri dan nilai tetangga elevasi, modul yang disebut makeVisualization secara otomatis menghitung legenda dinamis yang ditentukan oleh persentil 0,05 dan 0,95 dalam batasan tersebut. kotak. Selain itu, palet warna berbeda untuk membuat legenda peta tersedia di TAGEE: pelangi, inferno, cubehelix, merah2hijau, hijau2merah, ketinggian, dan aspek. Kode paket dan contoh minimal yang dapat direproduksi tersedia di https://github.com/zecojls/tagee (Bahan Tambahan).

5.4 EVALUASI STATISTIK

Kami melakukan evaluasi atribut TAGEE dengan membandingkan aspek dan kemiringan yang diperoleh dari dua fungsi GEE yang tersedia (ee.Terrain.aspect dan ee.Terrain.slope) dalam skala yang hampir global. Untuk tugas ini, kami menggunakan analisis korelasi Pearson dengan SRTM DEM 30m, yang berisi ketinggian dalam meter yang dibatasi pada area antara sekitar 60◦ lintang utara dan 56◦ lintang selatan. Penting untuk disebutkan bahwa untuk fungsi medan GEE yang tersedia saat ini, gradien lokal dihitung menggunakan empat tetangga yang terhubung dari setiap piksel, berbeda dari metode TAGEE yang diusulkan, yang menggunakan jendela piksel 3×3 dan juga mempertimbangkan geometri bola dalam perhitungannya. Dengan demikian, perbedaan minimal antar metode perhitungan diharapkan terjadi. Analisis ini dilakukan di GEE dan, selain korelasi Pearson, kami menghitung kesalahan absolut rata-rata relatif (MAE) antara output. MAE relatif diperkirakan dengan menghitung perbedaan absolut rata-rata antara dua raster dan menstandarkan hasilnya ke rentang (nilai maksimum dikurangi minimum) raster referensi.

Demikian pula, kami membandingkan hasil dari TAGEE dengan atribut medan yang dihitung oleh System for Automated Geoscienti fi c Analysis (SAGA) GIS versi 2.3.2. Dalam hal ini, kami mengunduh dari GEE DEM SRTM 30 m bersama dengan 12 atribut yang dihasilkan yang dihitung oleh TAGEE, semuanya meliputi Gunung Ararat (terletak antara 44,2◦ dan 44,5◦ E, dan 39,6◦ dan 39,8◦ N). Gunung Ararat dipilih karena variabilitas bentang alamnya yang tinggi dan ketersediaan peta yang dipublikasikan dari karya sebelumnya, yang memungkinkan perbandingan visual pola spasial. SRTM-DEM Gunung Ararat diproses di SAGA GIS menggunakan “Kemiringan, Aspek, Kelengkungan” dari modul Morfometri

Analisis Medan. Metode perhitungannya adalah “Evan (1979)” berdasarkan enam parameter dan polinomial orde 2, serupa dengan metode perhitungan TAGEE. Perbandingan dilakukan dengan menghitung koefisien korelasi Pearson (r) dan MAE relatif, dimana aspek, kemiringan, kelengkungan horizontal dan kelengkungan vertikal dari TAGEE dibandingkan dengan aspek, kemiringan, tangensial, dan kelengkungan profil dari SAGA GIS.

Hasil dan Pembahasan

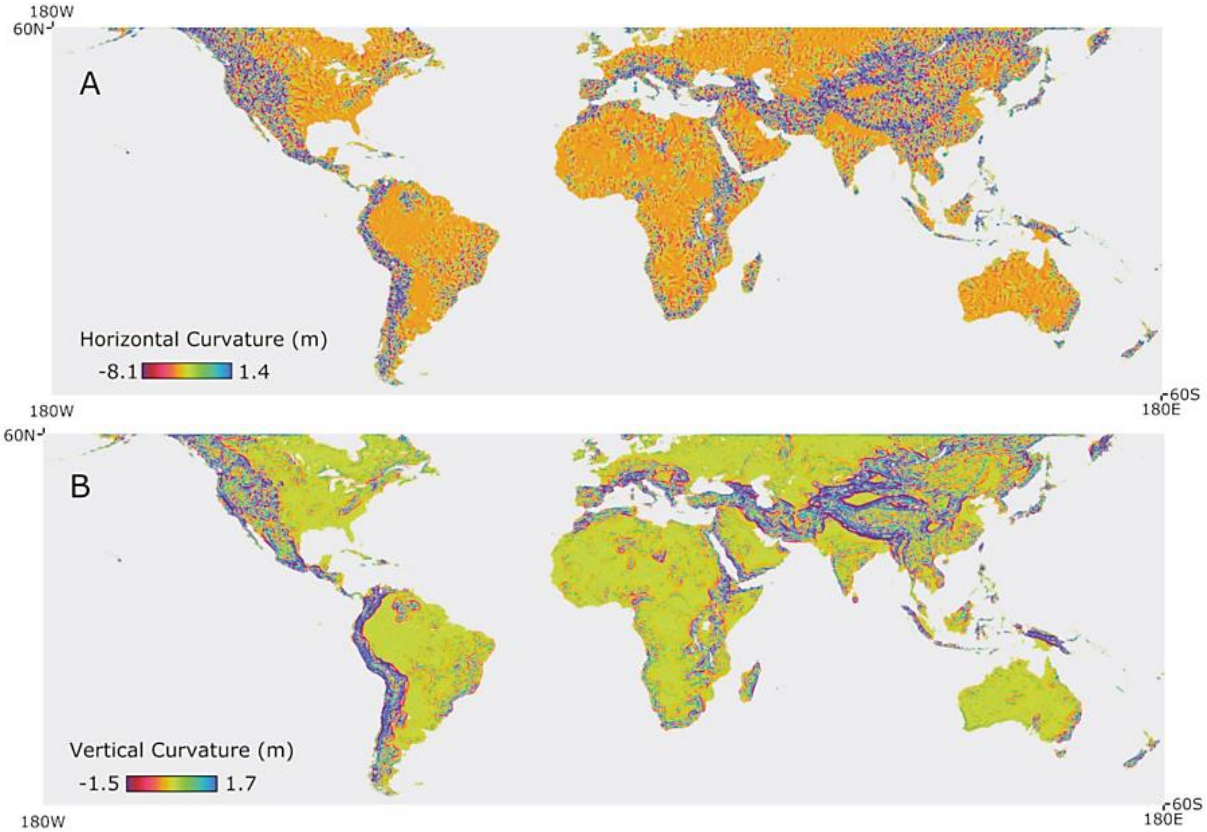
Analisis statistik mengungkapkan korelasi yang signifikan (p <0,01) dari keluaran TAGEE dengan atribut medan setara yang dihitung dari GEE dan SAGA GIS (Tabel 5.2). Kemiringan yang diperkirakan hampir mencapai tingkat global mencapai korelasi 0,98 (kesalahan 2%) antara TAGEE dan fungsi GEE, sedangkan aspek tersebut menghasilkan r Pearson sebesar 0,89 (kesalahan 13%). Korelasi aspek yang lebih rendah dapat dikaitkan dengan sifat dimensinya, yaitu variabel melingkar, serta perbedaan metode penghitungan antara TAGEE dan GEE. Meskipun perbedaannya kecil, TAGEE mengungkapkan pola spasial yang sama dan memungkinkan estimasi atribut tambahan pada skala global, seperti keutuhan utara, kelengkungan horizontal dan vertikal (masing-masing Gambar 5.3A–C). Pegunungan utama di Bumi, seperti Pegunungan Rocky di Amerika Utara, Andes di Amerika Selatan, Pegunungan Alpen di Eropa, Himalaya, dan dataran tinggi Tibet di Asia, dll., menyajikan kelengkungan tertinggi yang dihitung oleh TAGEE. Sebaliknya, dataran dan permukaan datar memiliki perkiraan terendah untuk kedua kelengkungan tersebut. Derajat orientasi ke Utara (Gambar 5.3A) juga menggambarkan bentang alam utama bumi.

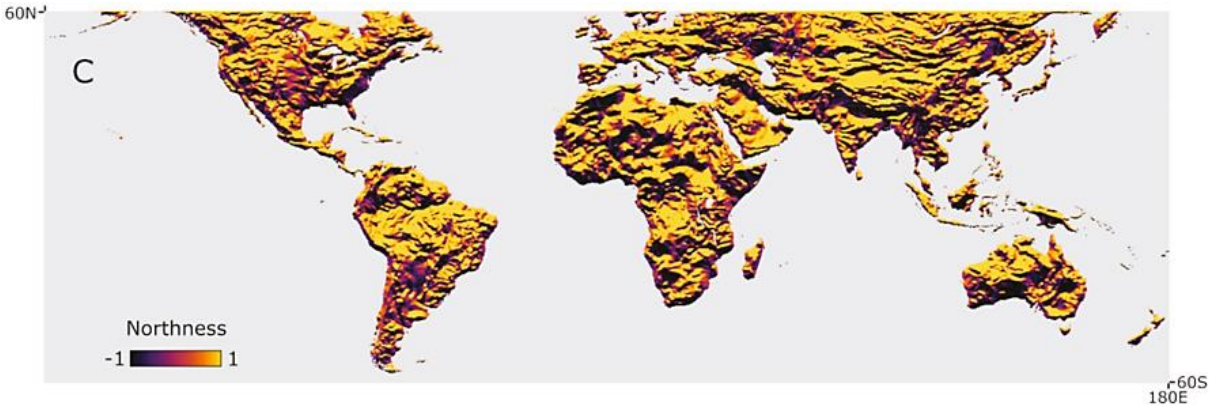
Tabel 5.2. Perbandingan atribut TAGEE dengan keluaran dari algoritma GEE dan SAGA GIS.

Atribut	Wilayah	Referensi	Pearson's r	rMAE[1]
Aspek	Near Global SRTM DEM 30m	GEE	0,89*	13%
Kemiringan	Near Global SRTM DEM 30m	GEE	0,98*	2%
Aspek	Mount Ararat SRTM DEM 30m	SAGA GIS	0,96*	4%
Kemiringan	Mount Ararat SRTM DEM 30m	SAGA GIS	0,98*	3%
Kelengkungan horizontal	Mount Ararat SRTM DEM 30m	SAGA GIS	0,98*	4%
Kelengkungan vertical	Mount Ararat SRTM DEM 30m	SAGA GIS	0,98*	4%

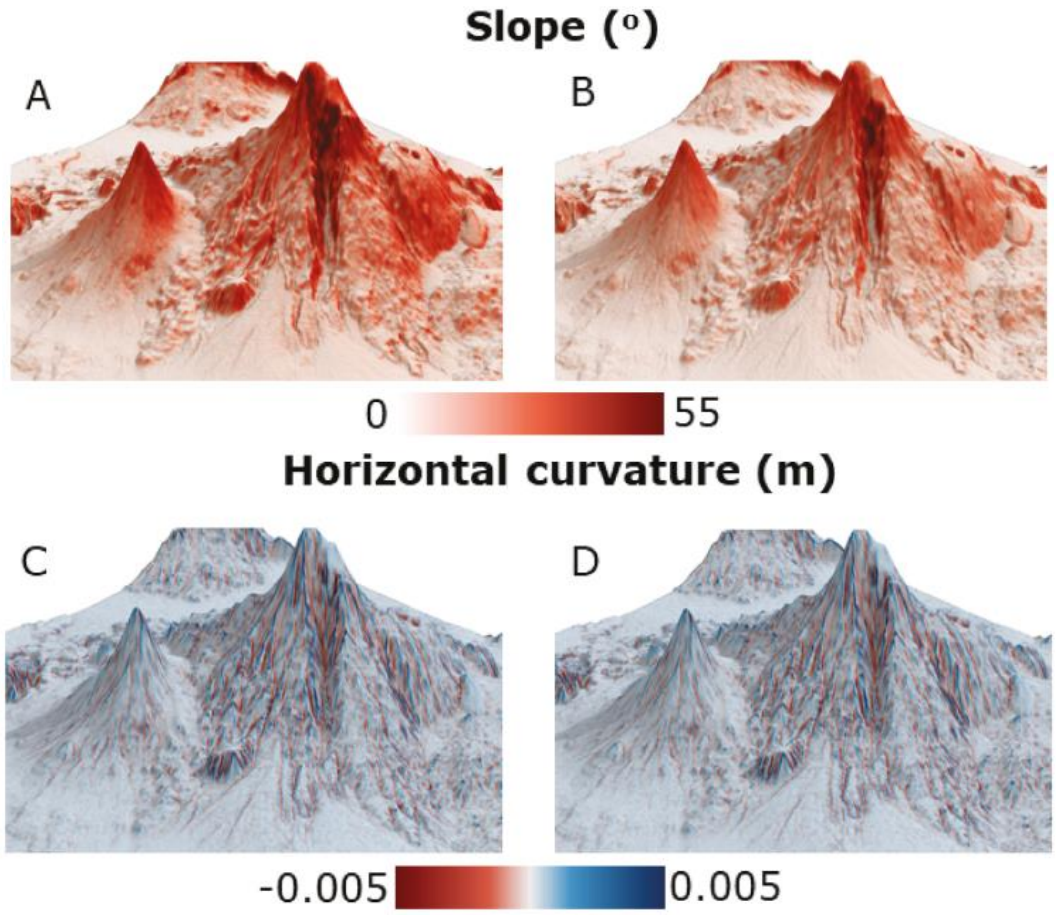
TAGEE dikembangkan di GEE untuk memanfaatkan komputasi platform berkinerja tinggi. Karena antarmuka berbasis cloud telah menciptakan banyak antusiasme dan keterlibatan dalam bidang penginderaan jauh dan geografis, banyak algoritma pemrosesan telah diadaptasi untuk membuat kemajuan substantif dalam tantangan global yang melibatkan pemrosesan data geografis berukuran besar. Dalam hal ini, GEE menyediakan citra penginderaan jauh berukuran petabyte yang tersedia untuk umum dan produk siap pakai lainnya. Pemrosesan paralel berkecepatan tinggi dari server GEE dan perpustakaan operator serta algoritme pembelajaran mesin yang tersedia oleh Antarmuka Pemrograman Aplikasi (API) dalam bahasa pengkodean populer, seperti JavaScript dan Python, memungkinkan pengguna menemukan, menganalisis, dan memvisualisasikan data besar geografis tanpa memerlukan akses ke superkomputer. Dalam kerangka kerja ini, TAGEE mendukung pengembangan analisis medan yang disesuaikan dengan data ketinggian berbeda di wilayah geografis yang luas.

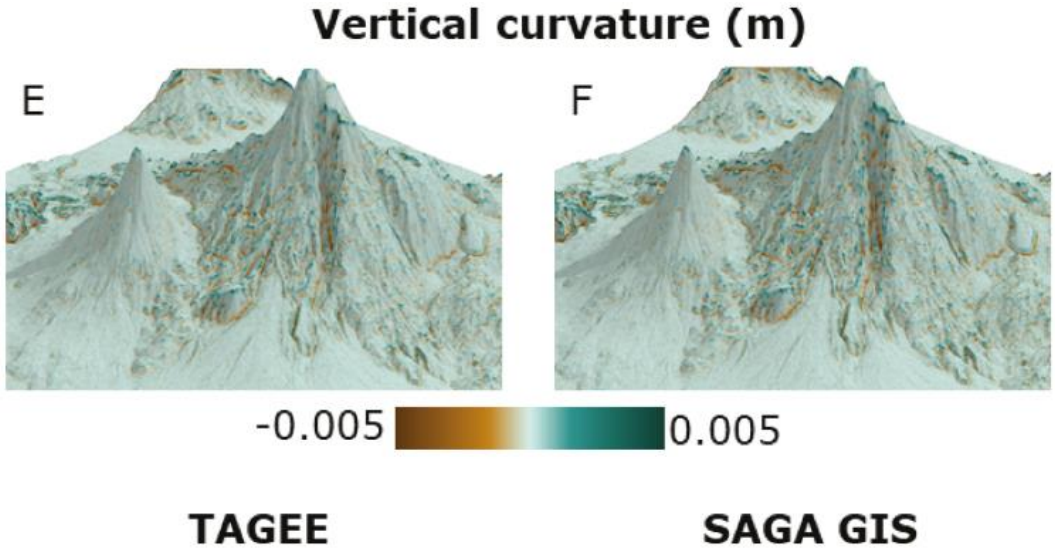
Ketika keluaran TAGEE dibandingkan dengan SAGA GIS (Tabel 5.2), evaluasi statistik menghasilkan korelasi yang signifikan dan tinggi untuk kemiringan, kelengkungan horizontal dan vertikal (Pearson's r sebesar 0,98, dengan selisih kesalahan sebesar 3 dan 4%).). Aspek dari TAGEE dan SAGA GIS memiliki koefisien korelasi yang lebih rendah, namun hasilnya lebih tinggi dibandingkan aspek dari algoritma GEE. Wilayah Gunung Ararat juga digunakan untuk membandingkan kemiringan, kelengkungan horizontal dan vertikal secara visual, yang dihitung dari TAGEE dan SAGA GIS (Gambar 5.4). Visualisasi 3D menunjukkan kemiripan yang tinggi antara kedua peta, namun beberapa perbedaan kecil dapat divisualisasikan melalui intensitas warna. Hal ini terjadi pada kemiringan Gunung Ararat yang dihitung dengan TAGEE (Gambar 5.4A), yang memiliki intensitas lebih tinggi dibandingkan kemiringan SAGA GIS (Gambar 5.4B). Intensitas kelengkungan vertikal yang dihitung dengan SAGA GIS sedikit lebih tinggi juga terlihat di tepi Gunung Ararat (Gambar 4F). Meskipun kecil, perbedaan visual ini menegaskan kesalahan relatif kedua metode (Tabel 5.2). Selain itu, pola spasial dari aspek, kemiringan, dan kelengkungan dari TAGEE menunjukkan korespondensi yang tinggi dengan peta medan Gunung Ararat yang tersedia memperkuat kepercayaan metode perhitungan TAGEE.





Gambar 5.3. Contoh atribut medan yang dihitung dari paket TAGEE dan DEM SRTM 1 detik busur, ditampilkan untuk cakupan hampir global pada tingkat visualisasi 3 (resolusi piksel ~20 km): kelengkungan horizontal (A), kelengkungan vertikal (B) , dan Keutaran (C).





Gambar 5.4. Visualisasi 3D atribut medan yang dihasilkan di dekat Gunung Ararat: kemiringan, kelengkungan horizontal dan vertikal dari TAGEE (masing-masing A,C,E) dan SAGA GIS (B,D,F). Peta 3D ditampilkan dengan perbesaran vertikal 2.

Dalam buku ini, algoritma TAGEE dikembangkan untuk mempertimbangkan geometri bola dalam metode perhitungannya. Pendekatan ini berbeda dari teknik yang tersedia dalam GIS tradisional, di mana TAGEE mempertimbangkan jarak lingkaran besar DEM yang ditentukan oleh posisi Lintang dan Bujur. Perangkat lunak GIS umum, seperti SAGA GIS, memerlukan proyeksi DEM untuk memastikan data ketinggian memiliki ukuran piksel yang sama. Beberapa peneliti terus menerapkan algoritma grid persegi pada DEM sudut sferoidal yang sama, yang dapat menyebabkan kesalahan komputasi besar dalam model variabel morfometrik. Kesalahan relatif yang kecil antara TAGEE dan GEE atau SAGA GIS dapat dikaitkan dengan perbedaan dalam metode penghitungannya.

Terakhir, beberapa keterbatasan TAGEE juga dapat diperhatikan. Hanya variabel morfometrik lokal yang dapat dihitung oleh paket, yang mencakup atribut fluks dan bentuk. Atribut non-lokal, seperti daerah tangkapan air tertentu, tidak diterapkan karena tidak adanya teori analitis umum, yang masih sedikit dikembangkan, dan karena proses rekursi yang masih menantang dalam GEE. Selain itu, metode baru tersedia untuk menangani masalah utama analisis medan, yang mencakup perkiraan DEM, generalisasi dan denoising, serta penghitungan variabel morfometrik.

5.5 RINGKASAN

Paket yang diusulkan (TAGEE) dapat menghitung atribut medan menggunakan platform GEE berkinerja tinggi dengan akurasi yang setara dengan GIS tradisional. Pendekatan penggunaan geometri spheroidal tidak memerlukan proyeksi masukan data ketinggian untuk perhitungan atribut medan. Perbandingan antara algoritma menunjukkan bahwa TAGEE memperkirakan kemiringan dan aspek medan yang serupa dengan fungsi GEE yang tersedia. Keuntungan TAGEE dibandingkan fungsi yang tersedia saat ini adalah keluaran tambahan dapat dihasilkan, seperti indeks kelengkungan dan bentuk, yang dapat berguna untuk

pemetaan lingkungan dan studi pemodelan. Selain itu, kesepakatan yang baik juga ditemukan ketika TAGEE dibandingkan dengan keluaran setara dari SAGA GIS, yang mencapai koefisien korelasi Pearson antara 0,96 dan 0,98, dan perbedaan antara 3–4%. Dengan demikian, TAGEE menjadi alat yang layak untuk membuat analisis medan data geografis besar, yang dapat disesuaikan dengan resolusi spasial apa pun dan ditingkatkan hingga skala global.

BAB 6
INTEGRASI ANALISIS GEOVISUAL DENGAN PEMBELAJARAN MESIN

Memahami pola pergerakan manusia merupakan hal yang sangat penting dalam perencanaan dan pengelolaan transportasi. Kami mengusulkan untuk mengkaji pola perjalanan angkutan umum yang kompleks melalui jaringan angkutan umum berskala besar, yang merupakan tantangan karena melibatkan ribuan penumpang angkutan umum dan data yang sangat besar dari berbagai sumber. Selain itu, representasi dan visualisasi pola perjalanan yang ditemukan secara efisien sulit dilakukan mengingat banyaknya jumlah perjalanan transit. Untuk mengatasi tantangan ini, penelitian ini memanfaatkan metode pembelajaran mesin canggih untuk mengidentifikasi pola mobilitas yang berubah-ubah terhadap waktu berdasarkan data kartu pintar dan data perkotaan lainnya. Pendekatan yang diusulkan memberikan solusi komprehensif untuk melakukan pra-proses, menganalisis, dan memvisualisasikan pola perjalanan angkutan umum yang kompleks. Pendekatan ini pertama-tama menggabungkan data kartu pintar dengan data perkotaan lainnya untuk merekonstruksi perjalanan angkutan umum asli. Kami menggunakan dua metode pembelajaran mesin, termasuk algoritme pengelompokan untuk mengekstraksi koridor transit guna mewakili koneksi mobilitas utama antara berbagai wilayah dan algoritme penyematan grafik untuk menemukan struktur komunitas mobilitas hierarkis. Kami juga merancang bentuk visualisasi multi-skala yang ringkas dan efektif untuk mewakili dinamika perilaku perjalanan yang ditemukan. Prototipe pemetaan berbasis web interaktif dikembangkan untuk mengintegrasikan metode pembelajaran mesin tingkat lanjut dengan visualisasi spesifik untuk mengkarakterisasi pola perilaku perjalanan angkutan umum dan untuk memungkinkan eksplorasi visual pola mobilitas angkutan umum pada skala dan resolusi berbeda dalam ruang dan waktu. Pendekatan yang diusulkan dievaluasi menggunakan data transit besar multi-sumber (misalnya, data kartu pintar, data jaringan transit, dan data lintasan bus) yang dikumpulkan di Kota Shenzhen, Tiongkok. Evaluasi prototipe kami menunjukkan bahwa pendekatan analisis visual yang diusulkan menawarkan solusi terukur dan efektif untuk menemukan pola perjalanan yang bermakna di wilayah metropolitan besar.

6.1 PENDAHULUAN

Pemantauan pergerakan manusia merupakan hal yang sangat penting dalam perencanaan dan manajemen transportasi. Untuk memfasilitasi perencanaan angkutan umum dan manajemen operasional, penting untuk memahami pola pergerakan angkutan umum melintasi ruang dan waktu. Untungnya, teknologi pengumpulan data geografis yang canggih saat ini, seperti sistem penentuan posisi global, pemetaan digital, sistem pembayaran tarif otomatis dengan kartu pintar, dan teknik komunikasi nirkabel, menghasilkan banyak data transit yang bervariasi secara spasial dan temporal sehingga menciptakan peluang untuk menemukan pergerakan yang bermakna dan signifikan. pola di wilayah metropolitan besar. Berbagai metode penambangan data telah dikembangkan untuk mengungkap pola perilaku

perjalanan transit berdasarkan kumpulan data geografis heterogen ini, termasuk pengelompokan untuk segmentasi penumpang, pemodelan bahaya untuk analisis loyalitas, metode rantai perjalanan untuk estimasi tujuan, dan pemodelan pilihan untuk analisis aktivitas penumpang.

Selama beberapa tahun terakhir, banyak penelitian telah dilakukan untuk mengeksplorasi pola perjalanan perkotaan menggunakan berbagai pemodelan dan pendekatan analitis berdasarkan data mobilitas manusia yang masif, seperti model keseimbangan rute berbasis optimasi untuk pengentasan kemacetan, analisis korelasi berbasis clustering untuk kesamaan mobilitas dan hubungan sosial, penemuan pola mobilitas tingkat rendah, dan eksplorasi fragmentasi sosial multi-skala. Dengan ketersediaan data mobilitas manusia yang sangat besar, teknik pembelajaran mesin telah memainkan peran yang semakin penting dalam memperoleh pemahaman mendalam tentang perilaku mobilitas manusia, mulai dari penambangan pola pergerakan, prediksi mobilitas, dan klasifikasi mode pergerakan, hingga penemuan dan prediksi gaya hidup.

Baru-baru ini, banyak upaya untuk memvisualisasikan data mobilitas manusia dalam jumlah besar, termasuk data ponsel, data pergerakan taksi, dan data media sosial telah dilaporkan. Beberapa sistem telah dikembangkan untuk melakukan analisis visual pada data kartu pintar, yang bertujuan untuk menemukan pola perjalanan yang menonjol untuk meningkatkan perencanaan dan pengelolaan angkutan umum. Upaya ini sebagian besar berfokus pada desain visualisasi baru dengan menggabungkan informasi perjalanan individu ke dalam bentuk visual yang ringkas. Dengan alat visualisasi ini, pengguna dapat menemukan dan menganalisis karakteristik perjalanan penting secara efisien. Namun demikian, sebagian besar metode ini berfokus pada visualisasi pola pergerakan spatio-temporal yang sederhana dan intuitif, seperti variasi aliran berdasarkan tempat, peta aliran antar area, atau peta aksesibilitas. Meskipun perencana transportasi umum dan manajer operasional perlu mengungkap pola pergerakan kompleks pada skala spatio-temporal yang berbeda, alat intuitif mereka tidak memadai karena didasarkan pada metode statistik sederhana. Hal ini memotivasi kami untuk menyelidiki kemungkinan penerapan teknik pembelajaran mesin untuk mengidentifikasi pola pergerakan angkutan umum tingkat tinggi dan kompleks yang mendukung perencanaan dan pengelolaan angkutan umum tingkat lanjut.

Dapat dikatakan bahwa visualisasi harus ditingkatkan dengan metode pembelajaran mesin yang canggih, mengingat besarnya ukuran dan kompleksitas data transit. Selama beberapa tahun terakhir, para peneliti telah mengembangkan alat analisis visual untuk mendukung eksplorasi interaktif pola pergerakan spatio-temporal menggunakan data mobilitas dalam jumlah besar. Di antara upaya ini, metode pembelajaran mesin telah digunakan untuk penemuan dan analisis pola. Misalnya, von Landesberger dkk. mengusulkan untuk mengintegrasikan pengelompokan spatio-temporal interaktif dan representasi grafik agregat untuk menemukan pola pergerakan perkotaan yang abstrak menggunakan media sosial dan data ponsel. Kami memilih untuk menemukan koridor angkutan umum dan komunitas mobilitas menggunakan dua algoritme pembelajaran mesin yang canggih karena keduanya menghasilkan pola mobilitas angkutan umum tingkat tinggi yang kompleks dan

representatif yang berguna untuk angkutan umum dan perencanaan kota. Selain itu, kami mengembangkan bentuk visualisasi interaktif khusus untuk memfasilitasi pemahaman tentang koridor dan struktur komunitas yang teridentifikasi. Kami berpendapat bahwa kombinasi pembelajaran mesin dan geovisualisasi bermanfaat untuk memperoleh pemahaman mendalam tentang pola mobilitas angkutan umum yang kompleks di wilayah metropolitan besar. Kami mengusulkan untuk mengkaji pola perjalanan angkutan umum tingkat tinggi dan kompleks menggunakan analisis visual pada jaringan angkutan umum berskala besar, yang merupakan tantangan karena melibatkan ribuan penumpang angkutan umum dan sejumlah besar data dari berbagai sumber. Selain itu, representasi dan visualisasi pola perjalanan yang ditemukan secara efisien juga merupakan tugas yang sulit mengingat jumlah perjalanan transit yang besar. Untuk mengatasi tantangan ini, penelitian ini memanfaatkan metode pembelajaran mesin canggih untuk mengidentifikasi pola mobilitas yang bervariasi terhadap waktu berdasarkan data besar angkutan umum multi-sumber. Kami juga merancang bentuk visualisasi multi-skala yang ringkas untuk mewakili dinamika perilaku perjalanan yang ditemukan. Prototipe berbasis web dikembangkan untuk menerapkan pendekatan analisis geovisual yang diusulkan dalam antarmuka grafis terintegrasi, yang memungkinkan analisis mendalam terhadap data angkutan massal multi-sumber. Kami mengevaluasi prototipe dengan data transit realistis yang dikumpulkan di Kota Shenzhen, Tiongkok. Studi kegunaan empiris kami menunjukkan bahwa pendekatan kami dapat menawarkan solusi terukur dan efektif untuk menemukan pola perjalanan yang bermakna di wilayah metropolitan besar.

Dalam buku ini, kami bertujuan untuk mengidentifikasi pola pergerakan angkutan umum yang bervariasi secara spasial dan temporal berdasarkan data angkutan umum yang sangat besar melalui jaringan angkutan umum yang besar. Kami memberikan kontribusi teknis berikut:

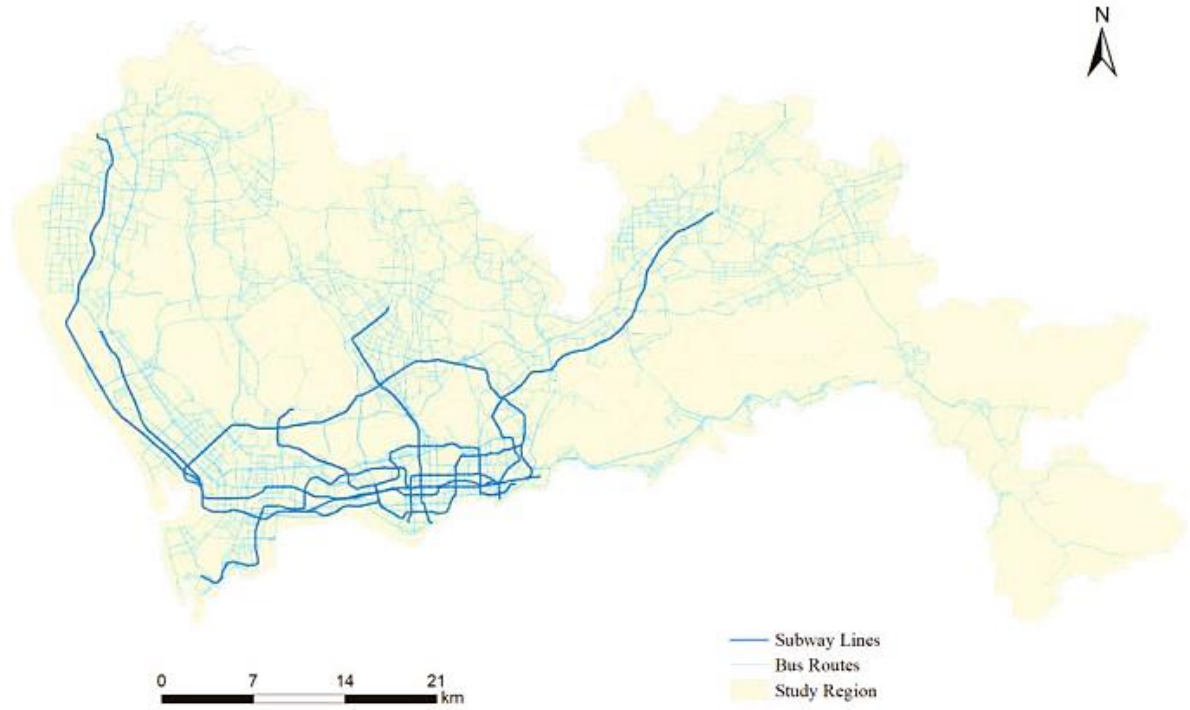
(1) Kami mengembangkan pendekatan analitik geovisual terintegrasi yang mengintegrasikan dua metode pembelajaran mesin canggih dengan peta interaktif untuk mengkarakterisasi dua jenis pola perilaku perjalanan transit tingkat tinggi yang kompleks, termasuk algoritma pengelompokan untuk mengidentifikasi koridor transit dan algoritma penyematan grafik untuk mengidentifikasi struktur komunitas mobilitas hierarkis.

(2) Kami merancang antarmuka analitik geovisual terpadu yang baru untuk pola pergerakan angkutan umum kompleks yang ditemukan, termasuk pandangan spesifik untuk memvisualisasikan komunitas dan koridor mobilitas yang teridentifikasi, sehingga memungkinkan pengguna biasa untuk memeriksa dan memahami pola yang selalu berubah ini pada skala dan perspektif berbeda.

6.2. KEBUTUHAN DATA

Dengan diterapkan pada kendaraan angkutan umum, sistem pembayaran tarif otomatis kartu pintar menyediakan cara yang efisien untuk mengumpulkan data perjalanan dalam jumlah besar di tingkat individu. Pendekatan yang diusulkan menggunakan data kartu

pintar (SCD) yang dikumpulkan di Kota Shenzhen, Tiongkok. Kota Shenzhen memiliki jaringan bus dan kereta bawah tanah besar yang terdiri dari 8 jalur kereta bawah tanah, 199 stasiun kereta bawah tanah, 808 rute bus, dan 6226 halte bus (Gambar 6.1).



Gambar 6.1. Wilayah studi dan jaringan angkutan umum.

Kami menggunakan SCD seminggu mulai tanggal 3 hingga 9 April 2017. SCD yang digunakan dalam penelitian ini menampung nama stasiun naik dan turun untuk setiap penumpang kereta bawah tanah. Penumpang bus tidak perlu mengetuk kartu pintarnya saat turun. Oleh karena itu, informasi turunnya halte tidak dicatat. Selain SCD, kami memiliki akses ke data lintasan bus, jaringan angkutan umum, dan data jaringan jalan raya. Ketiga dataset ini didaftarkan ke dalam kerangka georeferensi yang sama, yaitu sistem koordinat World Geodetic System 1984 (WGS 84) dan sistem koordinat Universal Transverse Mercator (UTM) Zona 50. Dataset jaringan angkutan umum berisi informasi lokasi, identifikasi, dan jadwal semua jalur kereta bawah tanah dan rute bus. Berdasarkan perangkat GPS yang dipasang di setiap bus, kami dapat memperoleh data lintasan bus seperti bujur dan lintang, kecepatan, dan arah perjalanan dengan interval kira-kira 20–60 detik. Selain itu, informasi identifikasi bus termasuk nomor plat, nomor jalur transit, dan nama perusahaan semuanya disimpan. Dengan lebih dari 6 juta catatan yang dikumpulkan setiap hari, ukuran SCD yang ditetapkan untuk minggu tersebut berjumlah 6,5 GB. Setiap hari, kumpulan data lintasan bus memiliki sekitar 63–73 juta catatan GPS.

Sistem angkutan umum terdiri dari beberapa komponen: halte bus, stasiun kereta bawah tanah, jalur bus dan kereta bawah tanah, kendaraan bus dan kereta bawah tanah, serta penumpang. Sebagian besar literatur yang ada berfokus pada analisis jalur/halte transit, jadwal, atau perjalanan transit gabungan. Beberapa pihak telah mengeksplorasi hubungan

antara perjalanan transit dan tempat menarik namun belum memanfaatkan metode pembelajaran mesin canggih untuk menganalisis pola perjalanan yang kompleks dan struktur mobilitas global. Mengingat banyaknya data perjalanan, seseorang mungkin ingin mengidentifikasi pola perjalanan spatio-temporal yang signifikan, mengungkap struktur mobilitas global, dan memvisualisasikannya pada peta interaktif. Misalnya, pertanyaan dapat diajukan untuk menemukan segmen jalan yang saling berhubungan yang memiliki pola permintaan perjalanan angkutan umum yang signifikan dalam skala global atau untuk menggambarkan wilayah dengan karakteristik perjalanan angkutan umum yang serupa. Pola spatio-temporal spesifik apa yang dapat ditemukan dari ruas dan kawasan jalan tersebut, dan bagaimana pola tersebut berkembang seiring berjalannya waktu? Bisakah kita bersama-sama mengkaji pola perjalanan angkutan umum dari berbagai aspek layanan angkutan umum dalam antarmuka pengguna interaktif yang terintegrasi?

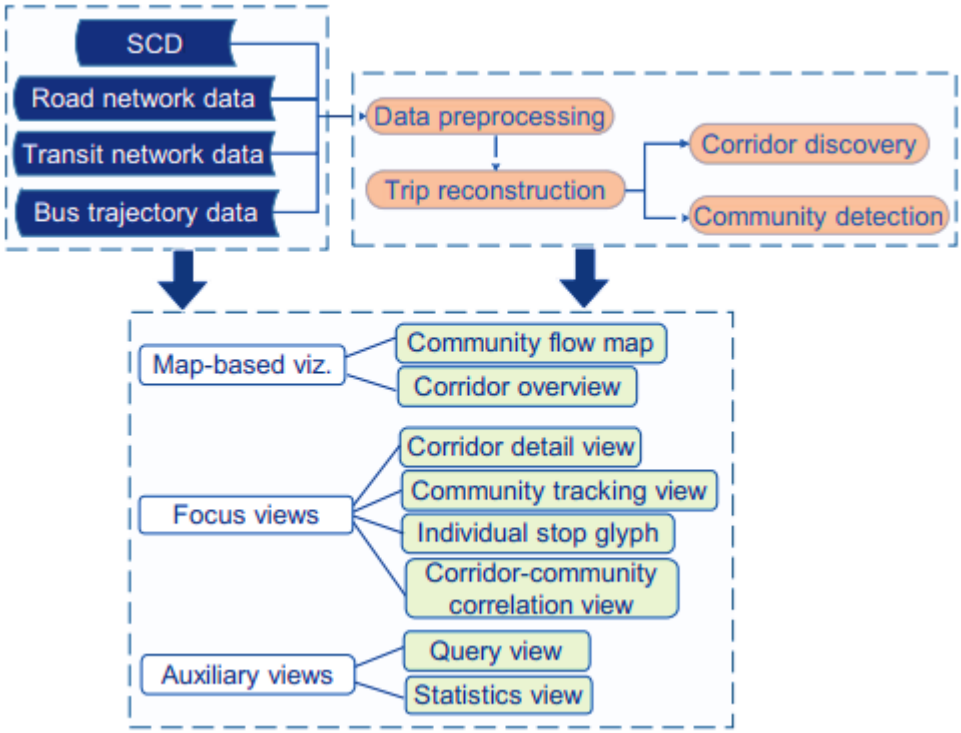
Untuk menjawab pertanyaan-pertanyaan ini, kita dapat mendefinisikan tugas analisis geovisual sebagai berikut:

1. Tugas penemuan pola global 1: menemukan struktur mobilitas hierarki berdasarkan data perjalanan transit dan menganalisis keterkaitannya;
2. Tugas penemuan pola global 2: mengidentifikasi koridor transit yang signifikan untuk interval waktu tertentu;
3. Eksplorasi pola lokal tugas 1: mengeksplorasi informasi intrinsik dari masing-masing koridor transit yang teridentifikasi;
4. Tugas eksplorasi pola lokal 2: mengkaji evolusi temporal dari struktur komunitas mobilitas yang ditemukan;
5. Tugas analisis komprehensif: merancang dan menerapkan pandangan terkait atau terintegrasi untuk menganalisis secara visual berbagai komponen layanan angkutan umum (termasuk koridor, struktur komunitas, dan halte) dan menemukan pola perjalanan.

Pendekatan yang diusulkan memberikan solusi komprehensif untuk melakukan pra-proses, menganalisis, dan memvisualisasikan pola perjalanan angkutan umum yang kompleks (Gambar 6.2). Pendekatan ini pertama-tama menggabungkan SCD dengan data perkotaan lainnya untuk merekonstruksi perjalanan angkutan umum. Kemudian mengelompokkan wilayah studi menjadi unit wilayah hierarkis (yaitu, komunitas mobilitas angkutan umum) berdasarkan fitur mobilitas perjalanan dan fitur lokal statis menggunakan penyematan grafik. Berdasarkan data perjalanan angkutan umum yang dipulihkan, kami mengembangkan algoritma berbasis pengelompokan untuk mengekstraksi koridor angkutan umum guna mewakili interaksi mobilitas antar wilayah yang berbeda. Berdasarkan koridor yang terdeteksi dan komunitas mobilitas, kami mengembangkan berbagai bentuk visualisasi untuk mewakili pola pergerakan transit ini pada peta dan tampilan lainnya. Prototipe pemetaan berbasis web interaktif juga dikembangkan untuk memungkinkan eksplorasi visual struktur mobilitas dalam ruang dan waktu. Bentuk visualisasi khusus dirancang dan diimplementasikan dalam prototipe berbasis web untuk memfasilitasi analisis data angkutan umum berukuran besar, termasuk visualisasi berbasis peta, tampilan fokus, dan tampilan tambahan. Struktur komunitas

mobilitas dan koridor transit yang ditemukan dapat divisualisasikan pada peta interaktif. Tampilan fokus terdiri dari empat jenis visualisasi:

1. tampilan detail koridor yang menampilkan informasi perjalanan rinci berdasarkan peta skema yang disederhanakan untuk setiap koridor yang dipilih;
2. pandangan pelacakan komunitas yang menyajikan perubahan yang terus berkembang pada komunitas tertentu sepanjang waktu;
3. mesin terbang perhentian yang menggambarkan informasi statistik semua perjalanan yang berasal, berakhir, atau melewati halte bus atau stasiun kereta bawah tanah tertentu; dan
4. pandangan korelasi koridor-komunitas yang menggambarkan korelasi spatio-temporal antara koridor transit dan komunitas mobilitas menggunakan plot koordinat paralel.



Gambar 6.2. Gambaran umum pendekatan analitik geovisual yang diusulkan. SCD—data kartu pintar.

6.3 PRA-PEMROSESAN DATA DAN REKONSTRUKSI PERJALANAN

Mengikuti prosedur yang dikembangkan di Zhang et al., kami melakukan pra-pemrosesan data untuk kumpulan data asli dan merekonstruksi perjalanan transit, yang digunakan dalam tugas analisis geovisual berikutnya. Bagi seorang penumpang, perjalanan angkutan umum terdiri dari beberapa jalur perjalanan yang terhubung secara berurutan dengan tujuan perjalanan tertentu. Pendekatan analisis visual kami didasarkan pada perjalanan, bukan perjalanan, karena perjalanan lebih baik dalam mengungkapkan tuntutan perjalanan dan pola perilaku yang realistis. Pada subbagian ini, kami menjelaskan secara singkat langkah-langkah pra-pemrosesan data dan rekonstruksi perjalanan.

Setelah menghapus catatan lintasan SCD dan bus yang salah, kami mengoreksi nama perhentian dan lokasi yang tidak konsisten antara kumpulan data heterogen berdasarkan metode yang dikembangkan. Semua kumpulan data diimpor ke database Microsoft SQL Server di mana indeks spasial dibuat berdasarkan data jaringan transit untuk mempercepat kueri data dan rekonstruksi perjalanan. SCD asli dibagi menjadi kumpulan data berbasis kereta bawah tanah dan berbasis bus pada setiap tanggal karena catatan berbasis kereta bawah tanah memiliki informasi naik dan turun secara lengkap dan prosedur estimasi pemberhentian turun diharapkan untuk SCD berbasis bus.

Pertama, kami perlu memperkirakan halte naik dan turun untuk perjalanan berbasis bus. Untuk setiap catatan SCD berbasis bus, halte keberangkatan dapat diidentifikasi dengan mencocokkan nomor pelat dari SCD dan kumpulan data lintasan bus untuk menemukan titik pengambilan sampel GPS yang mendekati waktu naik pesawat, yang kemudian dicocokkan dengan jaringan transit untuk menemukan kemungkinan besar pemberhentian boarding. Kemudian, kami melanjutkan untuk memperkirakan perhentian turun:

a. perhentian turun dapat dengan mudah diperoleh dengan mencari perhentian terdekat dengan perhentian naik berikutnya jika penumpang naik lagi pada hari yang sama;
b. jika perjalanan saat ini adalah perjalanan terakhir pada hari itu, pemberhentian pertama pada hari berikutnya digunakan untuk memperkirakan kemungkinan pemberhentian pada perjalanan terakhir tersebut;
c. jika tidak, kami dapat mencari tanggal lain atau penumpang serupa lainnya untuk membuat perkiraan. Dengan tersedianya halte naik dan turun, perjalanan bus yang lengkap dapat dipulihkan. Jalur perjalanan bus ini kemudian dihubungkan dengan jalur kereta bawah tanah untuk merekonstruksi perjalanan yang lengkap jika jalur perjalanan yang dilakukan oleh orang yang sama berada dalam ambang batas 30 menit.

6.4 MENGEKSTRAKSI KORIDOR TRANSIT

Konsep “koridor transit” telah diadopsi secara luas dan dimasukkan ke dalam praktik perencanaan dan pengelolaan dunia nyata. Kami mendefinisikan koridor sebagai segmen jalan linier berarah yang terdiri dari beberapa halte transit dengan jumlah penumpang yang signifikan. Perhatikan bahwa koridor mungkin mempunyai banyak cabang dan mungkin tumpang tindih dengan koridor lain. Berdasarkan data perjalanan angkutan umum yang sangat besar, koridor angkutan umum yang bervariasi terhadap waktu dapat diekstraksi untuk mewakili pola permintaan perjalanan yang paling signifikan dalam ruang dan waktu. Algoritme ekstraksi koridor dikembangkan berdasarkan perjalanan angkutan umum, yang masing-masing ditandai oleh satu tujuan perjalanan. Setiap perjalanan dapat terdiri dari beberapa bagian, dan setiap bagian berhubungan dengan satu transaksi kartu pintar. Kami mengusulkan algoritma pengelompokan aliran berbagi. Algoritme ini didasarkan pada konsep “akumulasi arus transit”, yang menghitung jumlah pemberhentian yang dilewati setiap penumpang setelah naik pesawat. Jika dua halte mempunyai “akumulasi arus transit” bersama yang besar, maka halte di hilir “dapat diakses oleh arus transit langsung” dari halte yang berdekatan sebelumnya. Dimulai dari perhentian dengan “akumulasi arus transit” dalam jumlah besar,

algoritme secara berulang mengevaluasi perhentian yang berdekatan sepanjang arah perjalanan. Apabila perhentian berikutnya ini dapat diakses langsung oleh arus transit dari perhentian saat ini, maka ruas jalan antara perhentian saat ini dan perhentian berikutnya akan dihubungkan dengan koridor saat ini. Setelah pertumbuhan awal koridor, dilakukan proses pemangkasan dan penggabungan untuk menghilangkan calon koridor yang pendek dan tidak signifikan. Dengan algoritma pengelompokan ini, koridor linier dapat ditemukan secara dinamis untuk interval waktu tertentu. Algoritmanya dapat dijelaskan dalam langkah-langkah berikut:

1. Pemodelan jaringan. Jaringan angkutan umum dapat dimodelkan sebagai grafik berarah dan dipetakan ke jaringan jalan G (V, E), dimana V menunjukkan himpunan persimpangan jalan Vr dan halte transit Vt (Vt telah diproyeksikan ke dalam ruas jalan), E menunjukkan ruas jalan antara persimpangan jalan dan halte transit. Kami mengekstrak sekumpulan kecil segmen terhubung Ec yang simpul ujungnya mempunyai "akumulasi arus transit" bersama yang besar dan mengidentifikasinya sebagai koridor transit.
2. Menghitung akumulasi arus transit. Untuk setiap node v di Vt, jumlah penumpang yang naik pada v atau sebelum dicatat sebagai nv. Untuk setiap penumpang, jumlah pemberhentian yang ia lewati setelah naik pesawat dicatat hingga ia keluar dari kendaraan. Kemudian untuk setiap v, bilangan ini digunakan sebagai "akumulasi arus transit" di(v).
3. Inisialisasi koridor. Kami memilih titik-titik dengan jumlah akumulasi arus transit yang signifikan sebagai benih untuk menumbuhkan koridor.
4. Perluasan koridor. Node benih ditumpuk ke dalam antrian prioritas, diberi peringkat berdasarkan akumulasi aliran transitnya. Yang memiliki at(v) tertinggi dikeluarkan dan digunakan sebagai benih awal s0 untuk memperluas koridor. Dari s0, algoritme mencari satu perhentian yang berdekatan, s1, yang memenuhi kriteria signifikan "arus transit akumulasi bersama" antara s0 dan s1. "Arus transit akumulasi bersama" didefinisikan sebagai sa(0→1) = [at(1)− at(0)] /at(0), yakni rasio perubahan akumulasi arus transit untuk dua pemberhentian/node yang berdekatan. Sedangkan kedua node tersebut harus memenuhi kriteria lain yaitu, "aliran transit bersama", yang didefinisikan sebagai st(0→1) = n0 n1. Jika kedua node memenuhi kedua kriteria tersebut, algoritme akan memperluas koridor dari node 0 ke 1. Prosedur ini berulang hingga tidak ada node hilir yang memenuhi kedua kriteria tersebut. Kemudian benih lain dalam antrian diambil untuk menumbuhkan koridor lain, sampai semua benih muncul.
5. Pemangkasan koridor. Kita perlu memangkas koridor-koridor yang pendek (kurang dari 4 perhentian) atau koridor-koridor yang tidak signifikan (arus angkutan umum kurang dari ambang batas yang telah ditentukan).
6. Penggabungan koridor. Langkah terakhir ini adalah menggabungkan koridor-koridor yang sudah terhubung atau tumpang tindih.

Biasanya, algoritme dapat mengekstrak 5–10 koridor transit untuk jam sibuk dan non-puncak selama hari kerja dan akhir pekan berdasarkan data perjalanan yang kami hasilkan.

Menemukan Komunitas Mobilitas

Sangat diharapkan untuk mewakili struktur mobilitas perkotaan tingkat tinggi dengan komunitas multi-skala ketika berhadapan dengan data mobilitas yang sangat banyak. Setiap komunitas mobilitas ditampilkan dengan karakteristik perjalanan yang serupa. Representasi struktur komunitas yang hierarkis secara signifikan dapat memudahkan pemahaman interkoneksi antarwilayah dalam suatu kota. Secara tradisional, pembangunan struktur komunitas menggunakan algoritma deteksi komunitas yang dikembangkan dalam ilmu jaringan. Algoritme deteksi komunitas ini pertama-tama membuat grafik untuk mewakili koneksi antar node dan kemudian menggunakan metode pengelompokan, optimasi, atau inferensi statistik untuk membagi keseluruhan grafik menjadi beberapa kelompok, memastikan node dalam setiap grup terhubung lebih padat dibandingkan node eksternal. Namun, penumpang angkutan umum biasanya melakukan perjalanan jauh dari tempat asal mereka, dan perilaku mobilitas ini harus diperhitungkan ketika mengambil struktur mobilitas angkutan umum. Studi ini mengusulkan definisi komunitas yang berbeda yang tidak hanya mempertimbangkan statistik perjalanan lokal tetapi juga tujuan perjalanan dan karakteristik perjalanan dinamis lainnya seperti frekuensi perjalanan dan pola perpindahan. Semua informasi ini dapat dengan mudah dihitung dari data perjalanan.

Daripada menandai setiap stasiun kereta bawah tanah atau halte bus dengan simpul grafik, kami menggunakan partisi wilayah dengan kemungkinan stasiun keberangkatan serupa di dekatnya. Pertama-tama kami mempartisi wilayah penelitian menjadi sel-sel grid biasa, yang masing-masing memiliki ukuran 100 m × 100 m. Kami menghapus sel-sel jaringan yang terletak di daerah pegunungan dan perairan (tidak dapat diakses oleh layanan transit). Kemudian vektor untuk setiap sel dibangun untuk mencatat kemungkinan stasiun keberangkatan (halte) yang dekat dengannya. Terakhir, algoritma heuristik digunakan untuk menggabungkan sel-sel tetangga dengan vektor yang paling mirip. Setelah penggabungan sel grid asli, diperoleh 18.109 grup grid, yang sebagian besar terdiri dari 2–7 sel grid asli. Kelompok grid ini kemudian dilambangkan sebagai node grafik, yang jumlahnya jauh lebih kecil dibandingkan perhentian transit asli, sehingga secara signifikan mengurangi biaya komputasi dalam pendeteksian komunitas.

Algoritme pendeteksian komunitas tradisional tidak dapat menangani masalah kami dan tidak dapat diskalakan ke jaringan angkutan umum yang besar. Untuk menangani fitur perilaku perjalanan yang kompleks, kami menggunakan metode penyematan grafik untuk mengungkap komunitas dinamis dari SCD yang realistis. Penyematan grafik bertujuan untuk menghasilkan representasi vektor yang kompak untuk setiap node dan mempertahankan struktur grafik dalam ruang berdimensi rendah.

Kita mendefinisikan graf berbobot berarah Gt(V, E) untuk selang waktu t. V adalah himpunan grup grid, dan E mewakili tepi koneksi transit antar node di V. Setiap tepi e diberi bobot oleh arus lalu lintas realistis antara node asal dan node akhir selama t. Berdasarkan bobot ini, kita dapat membuat matriks arus lalu lintas F, dimana $f_{i \to j}$ menyatakan jumlah penumpang transit yang melakukan perjalanan dari node i ke j. Kami juga membuat matriks ketetanggaan A untuk mendeskripsikan konektivitas lokal antar node grafik. Matriks A dapat

digunakan untuk merepresentasikan kedekatan konektivitas transit tingkat pertama. Struktur jaringan global dapat dipertahankan melalui kedekatan tingkat tinggi berdasarkan matriks arus lalu lintas F. Kedekatan tingkat tinggi didefinisikan sebagai kesamaan antara struktur konektivitas lalu lintas dari sepasang node.

Karena perilaku perjalanan angkutan umum sebagian besar bersifat non-linier dan tidak stasioner, kami memanfaatkan metode pembelajaran mendalam untuk mempelajari penyematan jaringan. Kerangka kerja auto-encoder klasik (struktur penyematan jaringan dalam, SDNE) diadopsi untuk mempelajari representasi jaringan laten. Kerangka kerja auto-encoder terdiri dari encoder dan decoder. Encoder berisi beberapa lapisan, yang masing-masing dapat didefinisikan sebagai (Persamaan 1)

$$\boldsymbol{z}^{(i)} = \sigma\left(\boldsymbol{W}^{(i)}\boldsymbol{z}^{(i-1)} + \boldsymbol{b}^{(i)}\right)$$

dimana $\boldsymbol{z}^{(i)}$ menunjukkan representasi tersembunyi untuk lapisan ke-i dan $\boldsymbol{z}^{0}$ adalah data masukan asli X, yang merupakan vektor berdimensi-n. $\boldsymbol{W}^{(i)}$ dan $\boldsymbol{b}^{(i)}$ adalah parameter yang dapat dipelajari. $\sigma(.)$ adalah fungsi aktivasi non-linier.

Jika kita menggunakan lapisan K di encoder, vektor masukan $\boldsymbol{z}^{0}$ akan dipetakan ke dalam representasi tersembunyi $\boldsymbol{z}^{(K)}$. Sejalan dengan itu, dekoder mengubah $\boldsymbol{z}^{(K)}$) kembali menjadi vektor Y yang direkonstruksi setelah melakukan operasi transformasi nonlinier lapisan K, (Persamaan 2)

$$\boldsymbol{z}'^{(j+1)} = \sigma\left(\boldsymbol{W}'^{(j)}\boldsymbol{z}'^{(j)} + \boldsymbol{b}'^{(j)}\right)$$

di mana $\boldsymbol{z}'^{(j)}$ menunjukkan vektor data yang direkonstruksi untuk lapisan ke-j dan $\boldsymbol{W}'^{(j)}$ dan $\boldsymbol{b}'^{(j)}$ adalah parameter yang dapat dipelajari. Perhatikan bahwa $\boldsymbol{z}'^{(0)} = \boldsymbol{z}'^{(K)}, \boldsymbol{Y} = \boldsymbol{z}'^{(K)}$

Parameter model dapat dipelajari dengan meminimalkan kesalahan rekonstruksi antara vektor $\boldsymbol{Y}$ yang direkonstruksi dan vektor masukan $\boldsymbol{X}$: (Persamaan 3)

$$L(X,Y) = \sum_{i=1}^{n} \|y_i - x_i\|_2^2$$

Jika beberapa fitur transit digunakan sebagai vektor input ke encoder (termasuk aliran, kecepatan, tujuan, dan frekuensi perjalanan), kita memperoleh $\boldsymbol{L_1}$. Jika matriks ketetanggaan $\boldsymbol{A}$ digunakan sebagai masukan, kita dapat membangun (persamaan 4)

$$L_2(A,Y) = \sum_{i,j>0}^{|V|} f_{i\to j}\left\|{y_j}^{(K)} - {y_i}^{(K)}\right\|_2^2$$

yang mempertahankan kedekatan tingkat tinggi G. Kedua fungsi kesalahan rekonstruksi dapat digabungkan secara linier menjadi fungsi kerugian gabungan yang komprehensif: (Persamaan 5)

$$L_{all} = L_1 + \alpha L_2$$

Model ini diinisialisasi dan dioptimalkan secara acak dengan penurunan gradien stokastik. Setelah konvergensi model, kita dapat memperoleh representasi penyematan akhir untuk semua node di G. Berdasarkan representasi node kompak yang dipelajari, kita dapat menggunakan pengelompokan hierarki untuk menghasilkan komunitas mobilitas hierarki.

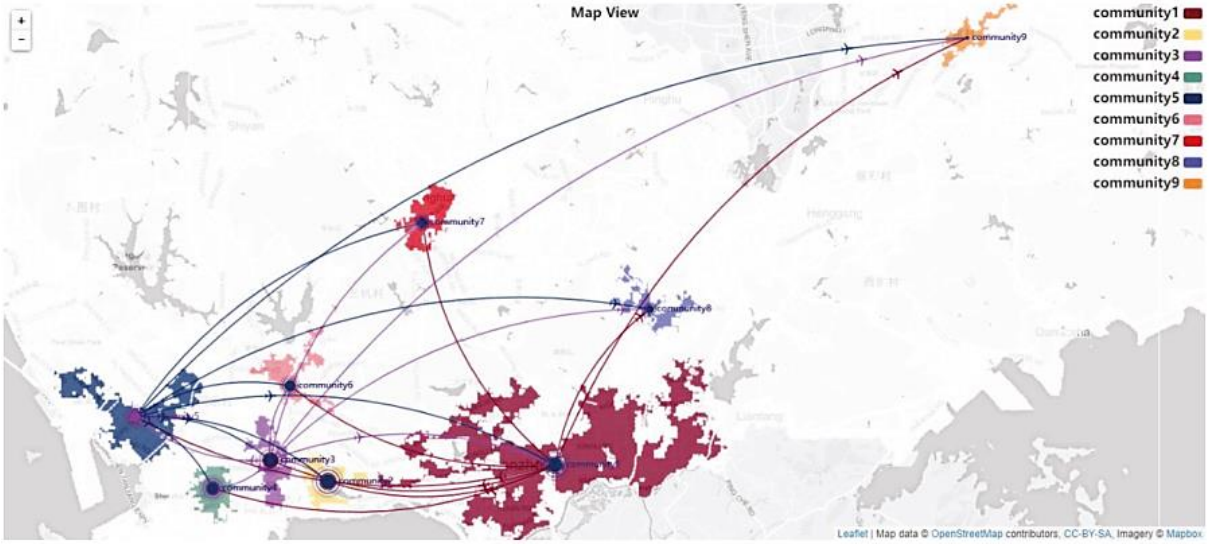
6.5 DESAIN ANALISIS VISUAL

Dalam literatur, studi visualisasi transportasi umum sebagian besar berfokus pada jaringan angkutan umum dan mewakili statistik perjalanan berdasarkan halte dan rute. Kami mengusulkan untuk mengkaji dan mengevaluasi layanan angkutan umum dari perspektif yang berbeda, yaitu komunitas mobilitas hierarkis dan koridor angkutan umum yang signifikan, selain dari jaringan angkutan umum. Beberapa desain visual diusulkan untuk memfasilitasi strategi analisis visual yang komprehensif ini.

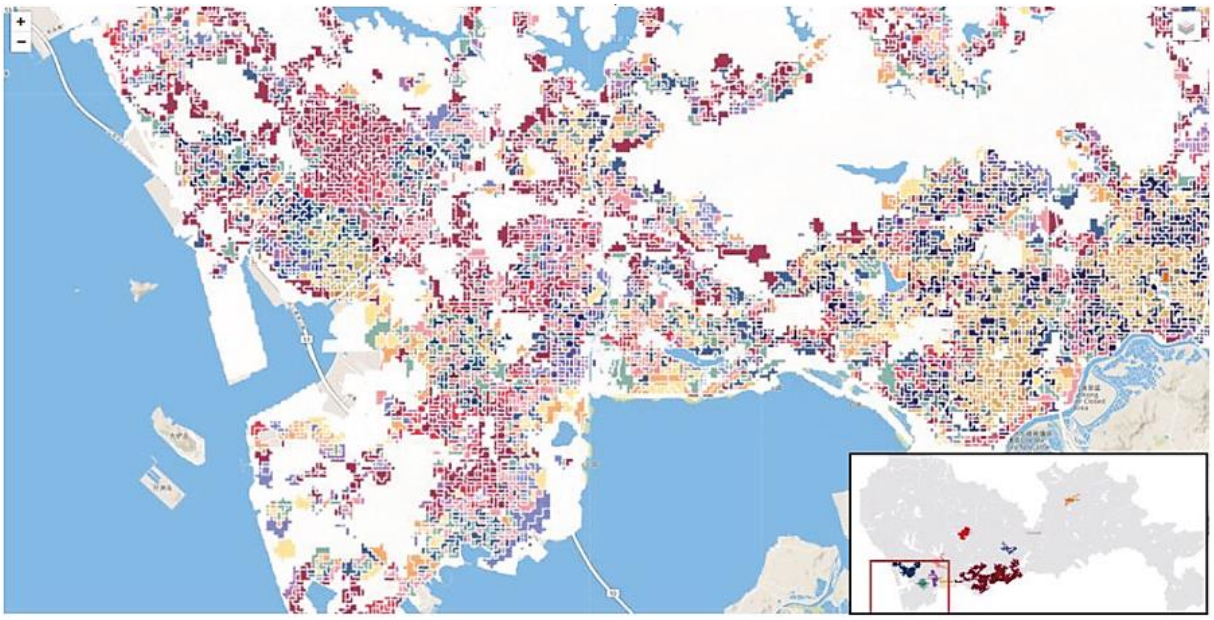
Desain visualisasi kami terdiri dari tiga jenis tampilan (Gambar 6.2): (1) Visualisasi berbasis peta yang menggunakan peta interaktif untuk menggambarkan komunitas mobilitas yang diekstraksi dan arus transit antar komunitas. Koridor yang terdeteksi juga diilustrasikan dalam tampilan peta. (2) Tampilan fokus dirancang untuk menyajikan informasi rinci tentang koridor yang dipilih pengguna, komunitas mobilitas transit, dan perhentian transit individu. Korelasi antara koridor dan komunitas juga dapat divisualisasikan. (3) Tampilan tambahan lainnya, termasuk tampilan kueri dan tampilan statistik. Tampilan kueri memungkinkan pemilihan data interaktif untuk analisis visual untuk interval waktu apa pun. Tampilan statistik menggunakan diagram statistik untuk menyajikan ringkasan informasi koridor.

Komunitas Mobilitas

Berdasarkan SCD yang realistis, kita dapat mengekstraksi struktur komunitas mobilitas dua tingkat di Kota Shenzhen. Setelah melakukan penyematan grafik untuk grup grid g, kita dapat melakukan pengelompokan hierarki berdasarkan grup grid ini untuk menghasilkan dua tingkat komunitas mobilitas. Komunitas tingkat rendah hanya didasarkan pada jarak antar vektor yang tertanam. Berdasarkan komunitas tingkat rendah, kami selanjutnya dapat menghasilkan komunitas tingkat tinggi dengan memperhitungkan kedekatan dan keterpaduan spasial menggunakan metode regionalisasi. Gambar 6.3a mengilustrasikan struktur komunitas mobilitas tingkat tinggi pada tanggal 3 April (hari libur). Arus transit antar komunitas tingkat tinggi ini dipetakan untuk menggambarkan gambaran umum arus transit gabungan di seluruh wilayah studi. Struktur komunitas tingkat tinggi disukai untuk penemuan dan analisis pola perjalanan global. Saat memperbesar ke tingkat rendah, struktur komunitas terperinci divisualisasikan dengan warna berbeda yang mewakili tipe cluster berbeda (Gambar 6.3b). Dengan peta komunitas interaktif yang disajikan dalam bentuk gambar, pengguna dapat melakukan tugas penemuan pola global 1 untuk mengidentifikasi struktur komunitas mobilitas hierarkis dan memvisualisasikan interaksi antar komunitas dengan mudah.



(a)

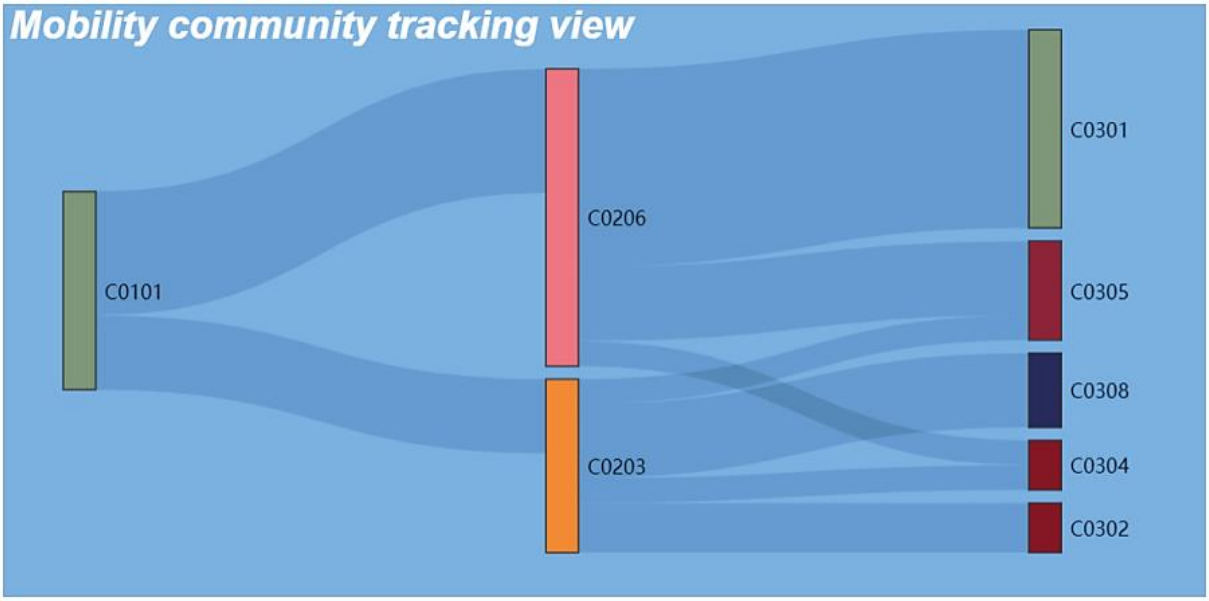


(b)

Gambar 6.3. Struktur masyarakat mobilitas berbasis transit pada hari libur. (a) Peta alur komunitas tingkat tinggi; (b) klaster komunitas tingkat rendah.

Dalam tampilan pelacakan komunitas mobilitas, seseorang dapat memilih (dari peta) satu komunitas mobilitas tertentu dan melacak perubahan temporal dalam bentuk dan aliran antara komunitas tersebut dan komunitas terdekat lainnya (Gambar 6.4). Seiring dengan berkembangnya struktur komunitas yang terdeteksi seiring berjalannya waktu, suatu komunitas mungkin mengalami perubahan yang berbeda-beda, termasuk terpecah menjadi komunitas terpisah atau bergabung dengan komunitas lain. Setiap komunitas diwakili oleh sebuah palang vertikal yang tingginya sebanding dengan jumlah perjalanan transit komunitas tersebut. Pita yang menghubungkan batang-batang di antara waktu yang berbeda mewakili arus transit antar komunitas. Pita lebar (sempit) menunjukkan volume perjalanan angkutan umum yang tinggi (rendah). Posisi vertikal bagian bawah batang juga menunjukkan hubungan

topologi komunitas dari tanggal yang berdekatan. Batang yang berjauhan satu sama lain menunjukkan komunitas yang juga berjauhan di peta. Saat kita memperbaiki posisi batang asli yang dipilih, hubungan yang tumpang tindih juga dapat terlihat dari posisi vertikal relatifnya antara dua batang dari tanggal yang berdekatan. Ketika struktur komunitas mengalami perubahan yang konstan, pengguna dapat melakukan tugas eksplorasi pola lokal 2 untuk melacak tren perkembangan komunitas yang dipilih dan mendapatkan pemahaman mendalam tentang struktur mobilitas di wilayah studi.



Gambar 6.4. Tampilan penelusuran mobilitas komunitas.

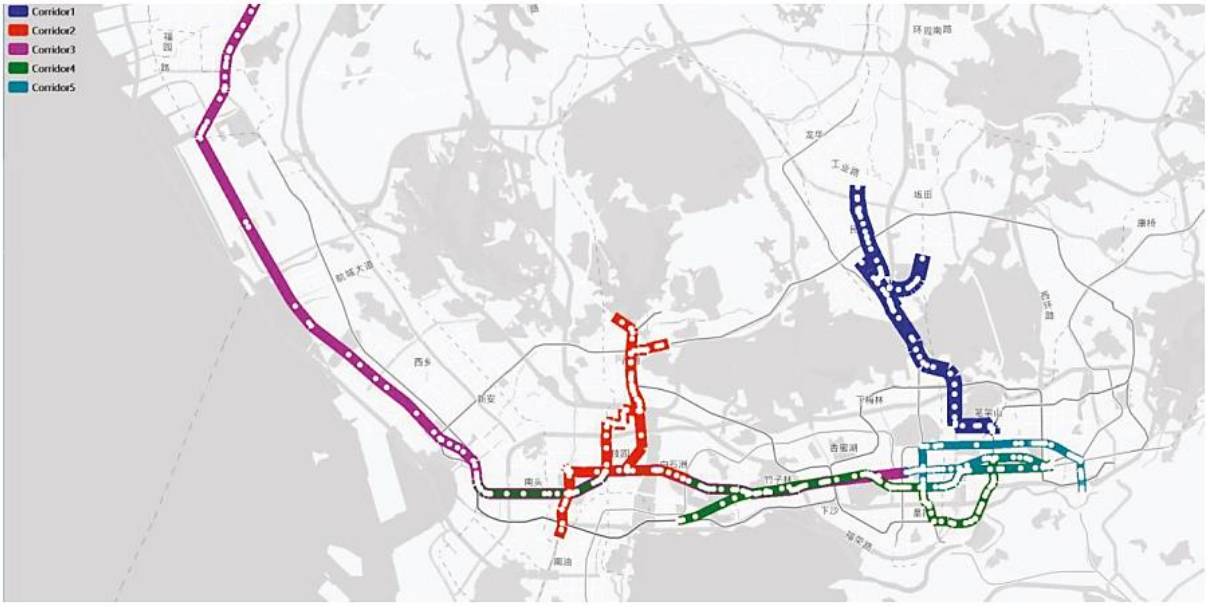
Kita dapat mengamati bahwa komunitas terpilih "C0101" (menunjukkan komunitas yang terdeteksi pada tanggal 1 April) memiliki jumlah perjalanan transit yang signifikan yang menghubungkan komunitas tersebut dan dua komunitas pada tanggal 2 April ("C0206" dan "C0203"). Keduanya selanjutnya terhubung dengan lima komunitas (C0301, C0302, C0304, C0305, dan C0308) pada tanggal 3 April. Kita dapat melihat bahwa "C0101" terhubung kuat dengan "C0206" dan "C0301", seperti yang ditunjukkan oleh lebar pita antara komunitas-komunitas ini.

Koridor

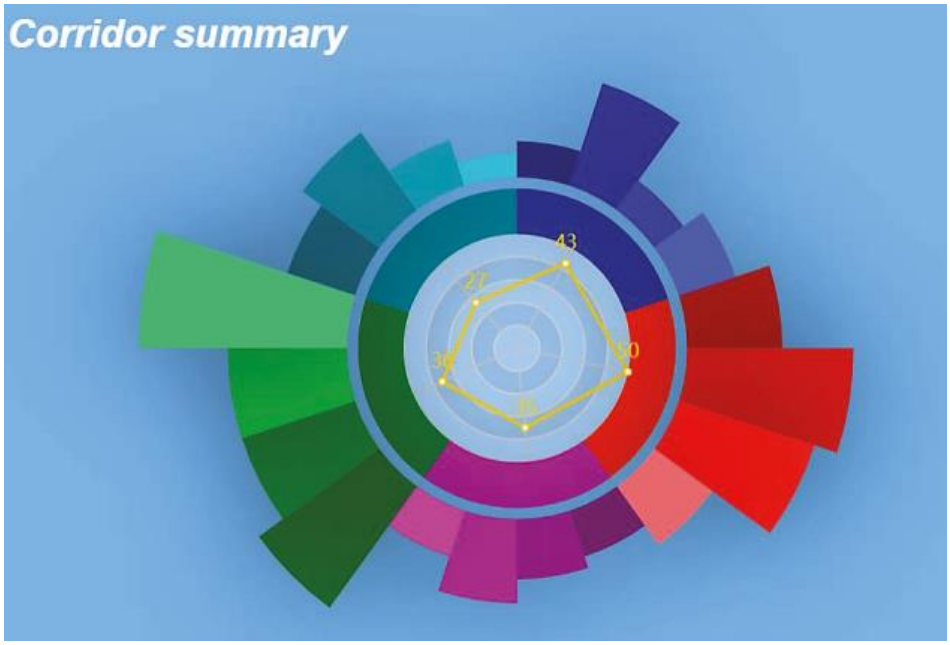
Gambar 6..5 menunjukkan lima koridor yang ditemukan pada hari kerja di peta. Lebar koridor mewakili ukuran arus transit. Arah aliran ditunjukkan oleh partikel animasi. Mesin terbang ringkasan khusus dalam tampilan statistik juga dirancang untuk menyajikan ringkasan ringkas semua koridor dalam tata letak radial, di mana setiap segmen berhubungan dengan koridor (Gambar 6.6). Untuk setiap koridor, jumlah penumpang yang naik dan turun dalam suatu koridor dibagi menjadi empat kategori: hanya asal perjalanan yang termasuk dalam koridor (tujuan berada di luar); hanya tujuan perjalanan yang terletak di dalam koridor (tidak asal); asal dan tujuan berada dalam koridor; dan kedua asal/tujuan berada di luar koridor. Empat batang digunakan untuk mewakili keempat jenis perjalanan ini. Ketinggian batang sebanding dengan jumlah perjalanan. Kinerja koridor secara keseluruhan juga diwakili oleh

garis titik-titik di lingkaran dalam. Kinerja dihitung sebagai rasio (persentase) waktu di dalam kendaraan versus waktu perjalanan keseluruhan dari perjalanan transit. Titik-titik yang dekat dengan pusat lingkaran menunjukkan kinerja yang rendah. Berdasarkan peta ikhtisar koridor ini, pengguna dapat melakukan tugas penemuan pola global 2 dan memeriksa distribusi koridor transit primer.

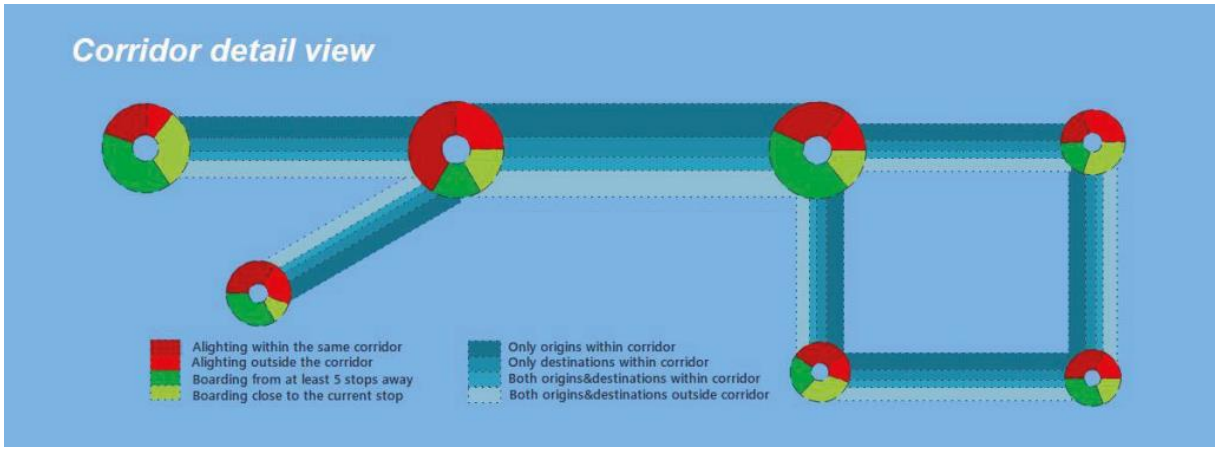
Pengguna dapat memilih dan mengamati detail koridor (Gambar 6.7). Tata letak koridor disederhanakan dalam bentuk skema untuk hanya mempertahankan koneksi topologi (mirip dengan peta metro). Empat jenis perjalanan yang disebutkan di atas divisualisasikan untuk koridor yang dipilih: setiap pita yang menghubungkan dua halte yang berdekatan dibagi menjadi empat komponen, dan lebarnya mewakili jumlah aliran. Perhentian utama dalam koridor digambarkan sebagai lingkaran, dengan warna merah/hijau mewakili jumlah naik/turun. Penumpang yang naik di halte selanjutnya dikategorikan menjadi dua kelompok: penumpang yang turun di koridor yang sama dan penumpang yang turun di luar koridor. Kedua kelompok tersebut masing-masing ditandai dengan warna merah tua dan merah terang. Demikian pula, penumpang yang turun di suatu halte dibagi menjadi dua kelompok: "naik dari jarak minimal 5 halte" dan "naik dekat dengan halte saat ini". Warna hijau tua dan hijau muda digunakan untuk menunjukkan kedua kelompok tersebut. Saat memilih koridor tertentu dan mengamati detailnya dalam tampilan detail koridor, pengguna dapat melakukan tugas eksplorasi pola lokal 1 untuk mengambil informasi naik, turun, asal, dan tujuan dalam bentuk visualisasi yang ringkas.



Gambar 6.5. Koridor angkutan umum pada jam sibuk pada hari kerja (08.00–10.00).



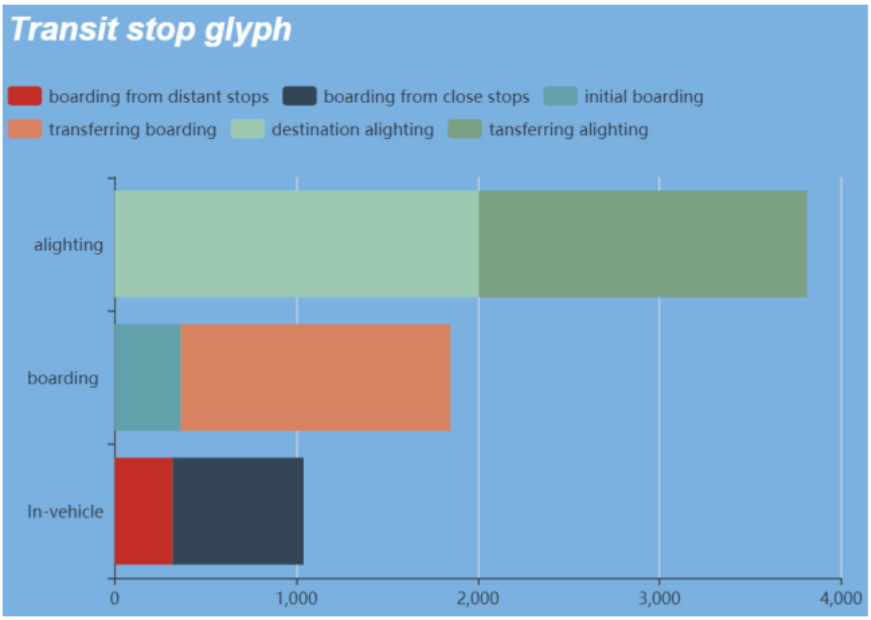
Gambar 6.6. Mesin terbang ringkasan koridor.



Gambar 6.7. Tampilan detail koridor.

Perhentian Transit

Informasi perjalanan dari masing-masing perhentian transit diplot dalam mesin terbang ketika pengguna mengklik perhentian di peta. Gambar 6.8 menunjukkan bahwa stop glyph dapat memvisualisasikan jumlah penumpang di dalam kendaraan, naik dan turun. Penumpang di dalam kendaraan dapat dibagi lagi menjadi naik dari halte jauh dan terdekat. Boarding penumpang terdiri dari boarding awal dan penumpang perpindahan. Penumpang yang turun terdiri dari mereka yang menyelesaikan perjalanannya dan mereka yang berpinda h di halte ini. Seseorang dapat dengan mudah mengetahui peran yang dimainkan oleh perhentian ini untuk keseluruhan jaringan transportasi umum: perhentian tersebut dapat berupa perhentian asal, perhentian tujuan, atau perhentian transfer.

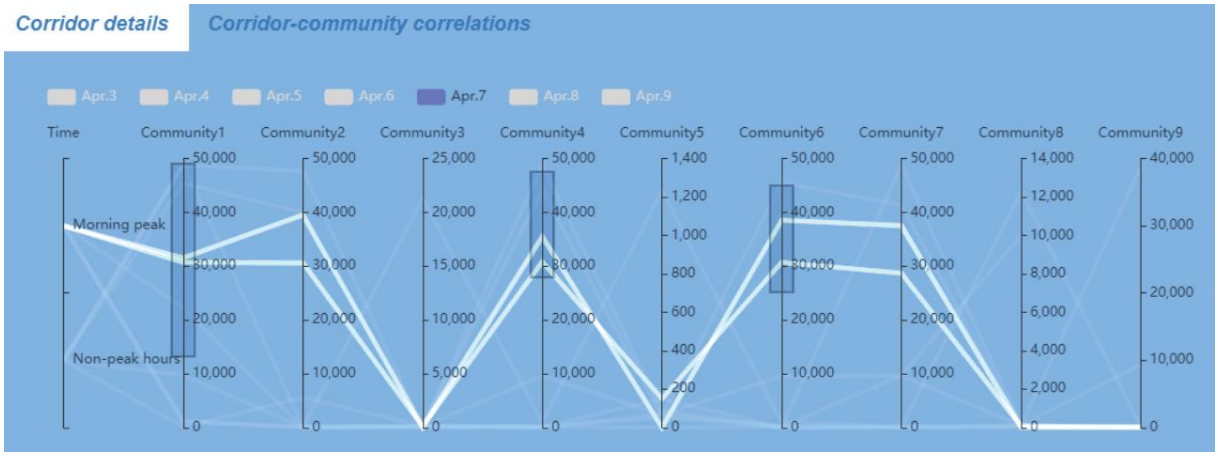


Gambar 6.8. Mesin terbang pemberhentian transit.

Korelasi antara Koridor dan Komunitas

Pada bagian sebelumnya, kami memperkenalkan pendekatan kami untuk menemukan koridor transit utama dengan kebutuhan perjalanan yang signifikan dan komunitas mobilitas dengan pola perjalanan serupa. Selain itu, analisis geovisual dapat dilakukan untuk menguji korelasi antara dua representasi mobilitas yang bervariasi terhadap waktu yang teridentifikasi ini. Plot koordinat paralel terintegrasi dirancang untuk menggambarkan korelasinya. Untuk interval waktu yang telah ditentukan sebelumnya, kita dapat menggambar koridor yang teridentifikasi sebagai polyline dan merepresentasikan komunitas yang ditemukan sebagai sumbu paralel vertikal. Untuk setiap koridor (yaitu polyline), titik potong pada suatu sumbu (yaitu komunitas) menunjukkan jumlah penumpang transit yang berasal dari komunitas menuju koridor tersebut. Dengan cara ini, korelasi spatio-temporal antara setiap koridor dan setiap komunitas dapat digambarkan secara kompak. Kita dapat dengan mudah menemukan komunitas mana yang menyumbang porsi terbesar arus transit ke suatu koridor tertentu atau mengidentifikasi koridor mana yang paling berkorelasi dengan komunitas tertentu. Kita juga bisa mengetahui komposisi koridor atau komunitas mana pun. Misalnya, plot koordinat paralel dapat mengungkapkan apakah sebagian besar perjalanan suatu komunitas berkorelasi dengan beberapa koridor atau tersebar merata di sejumlah koridor di seluruh kota. Perlu dicatat bahwa korelasi ini tidak setara dengan hubungan perpotongan antara koridor dan komunitas, yang terlihat jelas pada peta. Selama asal muasal perjalanan konstituen ke suatu koridor dapat ditelusuri ke suatu komunitas, maka koridor dan komunitas tersebut akan mempunyai korelasi. Jumlah korelasi ini mewakili intensitas interaksi antara pasangan koridor-komunitas. Untuk setiap sumbu vertikal, kotak penyaringan dapat ditentukan untuk menemukan koridor yang memenuhi kriteria rentang pencarian nomor aliran transit. Beberapa kotak penyaringan pada sumbu yang berbeda dapat dirancang secara interaktif untuk mengidentifikasi lebih jauh koridor-koridor yang berkorelasi dengan komunitas-

komunitas terpilih berdasarkan rentang arus transit tertentu (Gambar 6.9). Alat analisis geovisual ini memungkinkan pengguna melakukan tugas analisis komprehensif untuk menggali pengetahuan korelasi antara koridor dan komunitas mobilitas.



Gambar 6.9. Plot koordinat paralel untuk menggambarkan korelasi antara koridor transit dan mobilitas komunitas. Koridor yang ditemukan pada tanggal 7 April ditunjukkan pada gambar. Setelah menetapkan tiga kotak penyaringan, empat koridor dapat ditemukan dan disorot dalam plot.

6.6 IMPLEMENTASI DAN PROTOTIPE

Algoritma rekonstruksi perjalanan dan penemuan koridor diimplementasikan dalam C++. Algoritme ekstraksi koridor dan deteksi komunitas dilakukan pada komputer desktop dengan prosesor Intel™ Xeon E3-1240@3.70 GHz dan memori 16 GB, yang berjalan pada sistem operasi Microsoft Window 10.

Kami memilih perhentian yang memiliki jumlah arus lalu lintas persentil 85–90 sebagai benih koridor. Ambang batas “arus transit bersama” (st) ditetapkan sebagai persentil ke-50 dari jumlah arus transit. Ambang batas “akumulasi arus transit bersama” (sa) dapat diatur antara −15% dan 25%.

Struktur komunitas mobilitas diekstraksi menggunakan alat pengelompokan hierarki yang diimplementasikan dalam paket SciPy berdasarkan hasil penyematan jaringan yang dihasilkan oleh metode penyematan jaringan dalam struktur (SDNE). Algoritme penyematan grafik diimplementasikan menggunakan TensorFlow 1.14.0 dengan Python 3.6. Clustering dilakukan berdasarkan jarak Euclidean. Jaringan autoencoder berisi tiga lapisan: lapisan input berisi 18.109 neuron, yang sesuai dengan 18.109 grup grid di wilayah studi; lapisan tersembunyi memiliki 2000 neuron; dan lapisan keluaran menghasilkan vektor berdimensi 128 sebagai hasil akhir penyisipan untuk setiap node grafik. Lapisan yang lebih dalam akan menyebabkan penurunan kinerja, seperti yang ditunjukkan oleh pengujian kami. Inisialisasi parameter model didasarkan pada distribusi Gaussian (dengan mean $\mu = 1$ dan standar deviasi $\sigma = 0{,}01$). Bobot pada fungsi kerugian gabungan (Persamaan (5) ditetapkan sebesar 0,2 karena memberikan kinerja terbaik. Kecepatan pembelajaran ditetapkan sebesar 0,001. Untuk

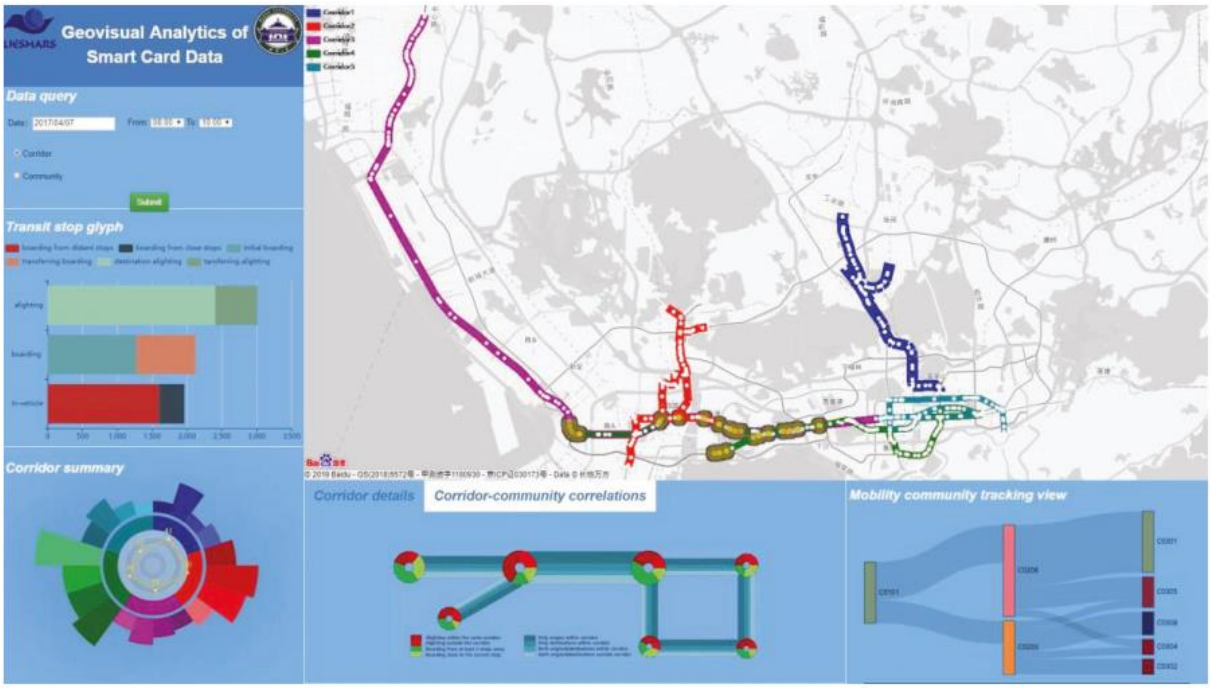
menghasilkan komunitas tingkat tinggi yang kohesif dan berdekatan, kami menerapkan algoritma regionalisasi, REDCAP, berdasarkan komunitas yang diproduksi oleh SciPy.

Kami membandingkan kinerja algoritma deteksi komunitas kami dengan algoritma deteksi komunitas klasik yang dikembangkan oleh Newman dan Leicht. Untuk mengevaluasi kinerja, kami menggunakan metrik modularitas yang diusulkan oleh Newman dan Girvan.

Seperti yang ditunjukkan oleh Tabel 6.1, algoritme penyematan grafik kami mengungguli algoritme Newman dan Leicht dengan selisih yang besar. Perhatikan bahwa studi kasus kami berbeda dari masalah deteksi komunitas biasa, di mana nilai modularitas yang lebih tinggi menunjukkan partisi komunitas yang baik. Karena kami mendorong masyarakat untuk melakukan perjalanan transit antarkomunitas yang padat dan perjalanan antarkomunitas yang jarang, maka nilai modularitas yang lebih rendah akan lebih baik.

Tabel 6.1. Perbandingan kinerja deteksi komunitas menggunakan metrik modularitas (komunitas tingkat rendah).

	Hari biasa	Weekend
Yang sudah ada	0.0372	0.0707
Yang dirancang	0.0218	0.0229



Gambar 6.10. Prototipe antarmuka pengguna berbasis web.

Desain visual diimplementasikan dalam prototipe berbasis web, yang dikembangkan dengan PyCharm Pro 2018.3.1 pada sistem operasi Windows 10. Modul visualisasi utama dikembangkan dalam JavaScript mengikuti standar HTML5 dan CSS3. Antarmuka pengguna terdiri dari empat komponen utama (Gambar 6.10): (1) tampilan kueri di bagian kiri atas; (2) tampilan peta di kanan atas; (3) tampilan statistik di pojok kiri bawah; dan (4) memfokuskan pandangan pada koridor dan komunitas di wilayah kanan bawah. Dalam tampilan kueri,

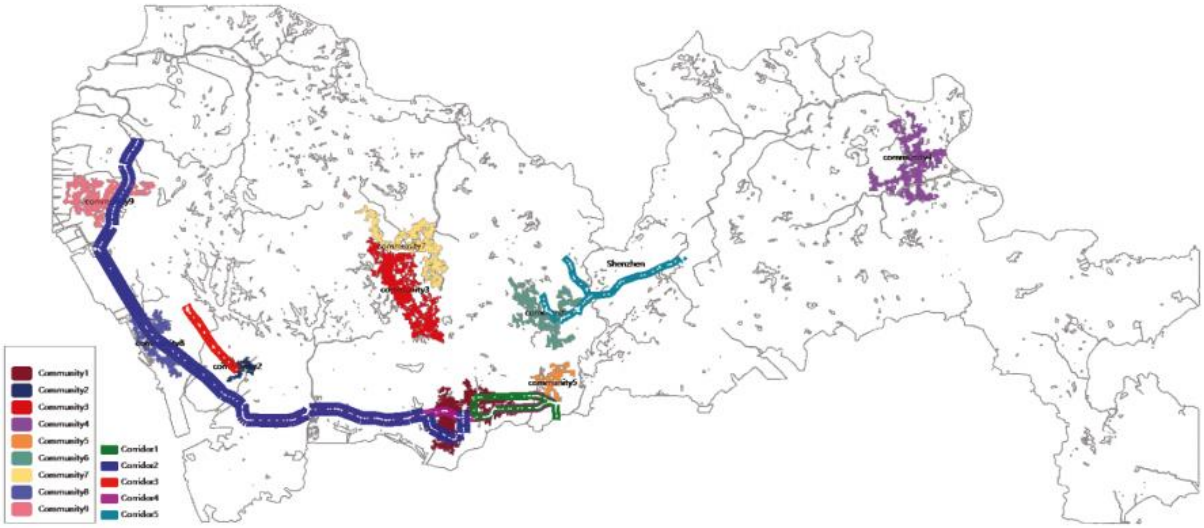
pengguna dapat menentukan rentang waktu dan memilih SCD yang termasuk dalam rentang ini untuk dianalisis. Tampilan peta menggambarkan koridor yang ditemukan dan struktur komunitas mobilitas. Koridor yang berbeda dibedakan berdasarkan warna yang berbeda. Tampilan peta menyematkan Peta Baidu sebagai peta latar belakang. Peta arus dapat dibuat untuk menggambarkan arus transit primer antara komunitas-komunitas besar. Tampilan fokus mengilustrasikan tiga jenis tampilan detail: tampilan detail koridor, tampilan pelacakan komunitas, dan tampilan korelasi koridor-komunitas. Semua pandangan ini terkait secara dinamis. Interaksi pengguna dalam tampilan mana pun akan berlaku untuk tampilan tertaut lainnya untuk data yang sama (komunitas, halte transit, atau koridor).

6.7 ALUR KERJA ANALISIS GEOVISUAL DAN CONTOHNYA

Biasanya, pengguna terlebih dahulu dapat menentukan rentang waktu SCD untuk analisis. Misalnya, dia dapat fokus pada jam sibuk pagi hari di hari kerja dan menggunakan algoritme back-end untuk mengekstrak struktur mobilitas dan koridor transit utama. Kemudian, baik koridor maupun struktur komunitas dapat divisualisasikan dalam berbagai tampilan yang saling terhubung untuk memungkinkan pemeriksaan lebih lanjut. Integrasi peta alur dan peta koridor dengan peta komunitas pada dua skala dapat membantu pengguna memahami struktur mobilitas transit kota secara keseluruhan. Sekilas, pengguna dapat mengidentifikasi daerah asal/tujuan utama dan berapa banyak penumpang yang melakukan perjalanan antardaerah tersebut. Sementara itu, tampilan statistik menyajikan informasi ringkasan seluruh koridor, yang memungkinkan pengguna membandingkan koridor yang diekstraksi dalam hal jenis perjalanan dan kinerjanya. Selain itu, pengguna dapat memilih koridor dan memvisualisasikannya dalam tampilan mendetail untuk mendapatkan informasi lebih lanjut tentang perjalanan yang ada di dalamnya. Pengguna juga dapat memilih perhentian transit mana pun untuk melihat penguraian perjalanan naik dan turun. Prototipe ini juga memungkinkan pengguna untuk memeriksa evolusi komunitas mobilitas yang dipilih dalam tampilan detail. Dengan semua tampilan terkait ini, pengguna dapat melakukan tugas analisis komprehensif untuk menemukan pola perjalanan transit global dan lokal di seluruh kota dari waktu ke waktu.

Misalnya, pengguna dapat menemukan komunitas mobilitas tingkat tinggi pada tanggal berapa pun. Gambar 6.3a menyajikan komunitas-komunitas ini untuk berlibur. Struktur komunitas yang teridentifikasi menyintesis pola perjalanan transit yang lebih mudah dipahami dibandingkan jejak transit massal yang asli. Komunitas terbesar (No. 1) terletak di sisi timur pusat kota, yang dilayani oleh beberapa jalur kereta bawah tanah dan puluhan stasiun kereta bawah tanah. Komunitas ini menarik banyak perjalanan berorientasi rekreasi yang berasal dari seluruh kota. Di sisi barat pusat kota, tiga komunitas terpisah (No. 2, 3, dan 4) dapat diamati, dan masing-masing menarik perjalanan jarak pendek di dekatnya. Komunitas lain tersebar di pinggiran kota, yang sebagian besar merupakan kawasan pemukiman. Banyak penumpang yang tinggal di komunitas pinggiran kota ini melakukan perjalanan ke pusat kota untuk tujuan rekreasi pada hari libur.

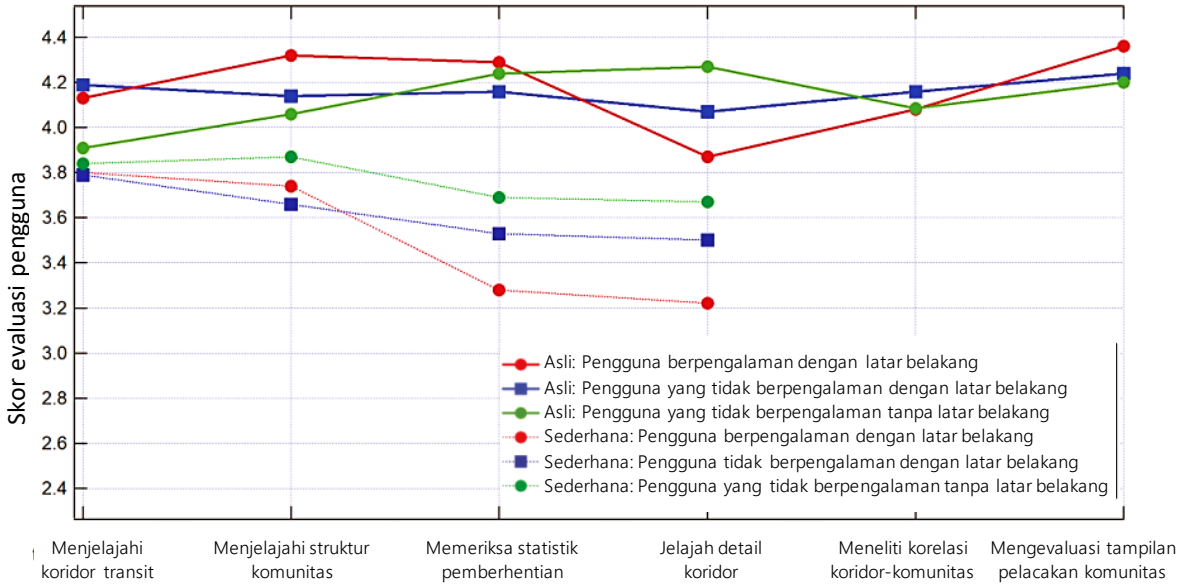
Gambar 6.11 menunjukkan bahwa mengkaji pola perjalanan intrinsik dengan mengintegrasikan koridor transit dengan komunitas mobilitas akan bermanfaat. Dapat diamati bahwa koridor paling menonjol menghubungkan komunitas No. 8 dan 9 serta komunitas No. 1, yang memiliki peluang kerja paling banyak di kota. Berdasarkan arah arus koridor (ditunjukkan oleh partikel animasi), terlihat bahwa banyak komuter yang melakukan perjalanan menuju komunitas No. 1 untuk bekerja pada pagi hari. Koridor lain di Timur Laut menunjukkan bahwa banyak penumpang yang tinggal di pinggiran terpencil melakukan perjalanan menuju komunitas No. 6, yang memiliki banyak kawasan industri dan perusahaan teknologi tinggi. Dengan peta interaktif ini, informasi koridor dan komunitas dapat digabungkan untuk menyelidiki lebih lanjut asal dan tujuan perjalanan untuk waktu dan tanggal yang berbeda, sehingga memperdalam pemahaman kita tentang pola pergerakan yang berkembang di seluruh kota. Peta-peta terpadu ini juga dapat berkontribusi pada penjelasan interaksi antara koridor transit dan komunitas mobilitas.



Gambar 6.11. Temuan koridor dan komunitas mobilitas pada pukul 11.00–13.00 pada hari kerja.

Dua puluh tiga pengguna diwawancarai untuk mendapatkan komentar dan umpan balik mengenai pendekatan analisis geovisual kami berdasarkan pengalaman mereka menggunakan prototipe. Enam belas di antaranya ahli di bidang transportasi umum, dan sembilan di antaranya memiliki pengetahuan pengembangan analisis geovisual (pengguna berpengalaman). Pengguna dapat diklasifikasikan menjadi tiga kelompok: (1) pengguna berpengalaman dengan latar belakang pengetahuan (9 pengguna); (2) pengguna belum berpengalaman dengan latar belakang pengetahuan (7 pengguna); dan (3) pengguna tidak berpengalaman tanpa latar belakang pengetahuan (7 pengguna). Sebelum mengizinkan mereka menggunakan prototipe, kami memperkenalkan usulan pendekatan analisis geovisual dan prototipe berbasis web. Kami meminta pengguna untuk mengevaluasi 6 tugas analisis geovisual: (1) menemukan dan memvisualisasikan koridor transit; (2) mengekstrak dan memvisualisasikan struktur komunitas mobilitas; (3) memperoleh informasi statistik ringkasan

pemberhentian transit; (4) mengevaluasi gambaran detail koridor; (5) mengevaluasi penelusuran mobilitas masyarakat; (6) untuk mengevaluasi pandangan korelasi koridor-komunitas. Skor numerik diperoleh dari kuesioner, dengan "0" menunjukkan pengalaman pengguna terburuk dan "5" menunjukkan yang terbaik. Gambar 6.12 merangkum skor pengguna yang diwawancarai.



Gambar 6.12. Skor evaluasi rata-rata untuk 6 tugas analisis geovisual yang dipilih.

Berdasarkan penilaian dan komentar, berbagai kelompok pengguna sepakat bahwa pendekatan analitik terintegrasi dan prototipe berbasis web kami memberikan solusi yang menarik dan dapat diterapkan untuk penemuan pola mobilitas manusia mengingat data angkutan umum yang sangat besar dan jaringan angkutan umum yang kompleks. Seperti yang ditunjukkan pada Gambar 6.12, pengguna berpengalaman dan pengguna dengan latar belakang pengetahuan cenderung memberikan peringkat lebih positif dibandingkan pengguna yang tidak berpengalaman atau pengguna tanpa latar belakang pengetahuan untuk sebagian besar tugas evaluasi. Mungkin diperlukan lebih banyak waktu bagi pengguna kelompok ketiga untuk memahami antarmuka dan fungsi sistem, sehingga mengurangi waktu mereka untuk sepenuhnya mengeksplorasi semua tampilan dan menyebabkan skor lebih rendah. Tugas evaluasi (1) dan (4) memiliki peringkat yang relatif rendah, mungkin karena konsep koridor transit yang masih asing bagi sebagian pengguna. Pengguna mungkin mengalami kesulitan dalam memilih dan memeriksa koridor tertentu di antara pandangan-pandangan yang berbeda, hal ini dikonfirmasi oleh wawancara umpan balik berikutnya. Tampilan detail koridor juga tidak intuitif untuk digunakan, menurut komentar pengguna, karena mengharuskan pengguna untuk sering mengalihkan fokus antara tampilan peta dan tampilan detail.

Kami juga menerapkan versi berbasis web yang disederhanakan untuk evaluasi pengguna. Dibandingkan dengan versi aslinya, prototipe yang disederhanakan hanya memiliki tampilan peta terintegrasi untuk menunjukkan koridor yang ditemukan dan struktur komunitas. Ini tidak mengimplementasikan tampilan tertaut dan hanya memiliki visualisasi terbatas (misalnya, tanpa partikel animasi untuk menunjukkan arah aliran di koridor, tanpa mesin terbang perhentian transit individu dan tampilan detail koridor).

Dengan sistem yang disederhanakan, kami melakukan wawancara evaluasi pengguna dengan tiga kelompok pengguna yang sama. Prosedur wawancara evaluasi yang sama dilakukan untuk meminta skor dan umpan balik mereka. Perlu dicatat bahwa hanya empat tugas evaluasi yang dievaluasi, yaitu eksplorasi koridor transit, eksplorasi struktur komunitas, pemeriksaan statistik pemberhentian, dan eksplorasi detail koridor. Skor evaluasi rata-rata untuk sistem yang disederhanakan juga ditunjukkan pada Gambar 12. Seperti yang dapat kita lihat, skor evaluasi ini jauh lebih rendah daripada skor yang diperoleh berdasarkan prototipe asli untuk empat tugas yang dievaluasi.

Dalam penelitian ini mengadopsi konsep struktur komunitas dan koridor transit untuk membangun pengetahuan mobilitas agregat tingkat tinggi dari SCD yang masif dan data perkotaan lainnya. Hasil dari algoritma deteksi komunitas dan penemuan koridor diintegrasikan ke dalam antarmuka visualisasi interaktif yang terdiri dari beberapa tampilan terkait untuk memungkinkan analisis visual yang efisien terhadap pola mobilitas transit spatio-temporal pada berbagai skala dan resolusi. Tampilan peta menawarkan antarmuka ikhtisar untuk membantu pengguna menyimpan informasi konteks ketika mereka fokus pada visualisasi koridor, komunitas, atau perhentian tertentu. Tampilan spesifik seperti tampilan detail koridor, tampilan pelacakan komunitas mobilitas, dan plot korelasi koordinat paralel, serta mesin terbang ringkasan (termasuk mesin terbang perhentian transit dan mesin terbang ringkasan koridor) melengkapi tampilan peta untuk menyediakan alat analisis geovisual yang intuitif untuk penemuan detail pengetahuan tentang komponen spesifik apa pun dari sistem angkutan umum. Keuntungan mengintegrasikan pembelajaran mesin dengan visualisasi interaktif dapat diringkas sebagai berikut:

(1) Ini menawarkan metode yang efisien dan efektif untuk mengeksplorasi sejumlah besar perjalanan transit, yang sebaliknya sulit untuk dianalisis dan divisualisasikan. Berdasarkan koridor yang ditemukan dan komunitas mobilitas, kita dapat fokus pada pola perjalanan yang paling signifikan sambil tetap memiliki kemampuan untuk mengeksplorasi detail perhentian mana pun.
(2) Ini memberikan antarmuka pengguna yang intuitif untuk menggabungkan berbagai tampilan yang memungkinkan pengguna biasa menganalisis perilaku perjalanan angkutan umum yang kompleks dari perspektif yang berbeda. Misalnya, koridor menghadirkan representasi tingkat tinggi dari perjalanan terkonsentrasi berdasarkan jaringan jalan raya, sedangkan komunitas mobilitas dihasilkan untuk mensintesis karakteristik perjalanan serupa di seluruh partisi wilayah studi.
(3) Hal ini bermanfaat bagi banyak aplikasi pengelolaan angkutan umum, seperti pemodelan permintaan, perencanaan angkutan umum, dan operasional sehari-hari,

karena keduanya memberikan pendekatan yang dapat diterapkan untuk menyoroti pola pergerakan agregat pada berbagai resolusi spasial dan temporal. Prototipe ini juga dapat digunakan oleh penumpang reguler untuk merencanakan perjalanan transit dan memilih tempat tinggal atau tempat kerja.

Bagi sebagian besar penduduk kota, perjalanan transit mengikuti ritme mingguan: mereka berangkat kerja setiap hari kerja dan menikmati waktu luang di akhir pekan. Data satu minggu mungkin cukup untuk mengekstraksi pola pergerakan angkutan umum di wilayah studi. Dalam literatur, kami juga menemukan peneliti lain yang juga menggunakan data angkutan umum selama satu minggu (yaitu data kartu pintar) untuk penelitian mereka. Misalnya, Long dan Thill meneliti hubungan pekerjaan-perumahan di Beijing dengan SCD berbasis bus satu minggu. Alsger dkk. memvalidasi algoritma estimasi asal-tujuan berdasarkan SCD satu minggu di Queensland Tenggara, Australia. Jika kita dapat mengakses data lintasan SCD dan GPS dari periode waktu lain, pendekatan analisis geovisual yang sama dapat segera diterapkan.

6.8 RINGKASAN

Dalam bab ini, kami menerapkan dua metode pembelajaran mesin, termasuk algoritme pengelompokan untuk mengekstraksi koridor transit dan algoritme penyematan grafik untuk menemukan struktur komunitas mobilitas. Representasi tingkat tinggi ini divisualisasikan dalam antarmuka interaktif berbasis web untuk memungkinkan pengguna memeriksa SCD masif dengan cara yang sangat teragregasi dan efisien. Prototipe kami menunjukkan bahwa pendekatan analisis visual yang diusulkan dapat menawarkan solusi terukur dan efektif untuk menemukan pola perjalanan yang bermakna di wilayah metropolitan besar. Kami berencana untuk meningkatkan kegunaan prototipe berdasarkan komentar pengguna dalam waktu dekat. Memungkinkan pengguna untuk menentukan konfigurasi algoritma dalam antarmuka pengguna grafis, yang berkontribusi pada pemahaman yang lebih baik tentang algoritma pembelajaran mesin yang mendasari, dan ini akan diimplementasikan dalam waktu dekat.

BAB 7
BIG DATA MINING MEDIA SOSIAL DAN ANALISIS SPATIO-TEMPORAL

Media sosial memuat banyak informasi geografis dan telah menjadi salah satu sumber data penting untuk mitigasi bahaya. Dibandingkan dengan metode pengumpulan informasi geografis terkait bencana yang tradisional, media sosial memiliki karakteristik penyediaan informasi secara real-time dan berbiaya rendah. Karena perkembangan teknologi penambangan data besar, kini lebih mudah untuk mengekstrak informasi geografis yang berguna terkait bencana dari data besar media sosial. Selain itu, banyak peneliti telah menggunakan teknologi terkait untuk mempelajari media sosial untuk mitigasi bencana. Namun, hanya sedikit peneliti yang menganggap ekstraksi emosi masyarakat (terutama emosi yang sangat mendalam) sebagai atribut informasi geografis terkait bencana untuk membantu mitigasi bencana. Dikombinasikan dengan kemampuan analisis spatio-temporal yang kuat dari sistem informasi geografis (GIS), informasi emosional masyarakat yang terkandung dalam media sosial dapat membantu kita memahami bencana secara lebih rinci dibandingkan dengan yang dapat diperoleh dari metode tradisional. Namun, data media sosial cukup kompleks dan terfragmentasi, baik dari segi format maupun semantik, khususnya untuk media sosial Tiongkok. Oleh karena itu, diperlukan algoritma yang lebih efisien. Dalam bab ini, kami mempertimbangkan gempa bumi yang terjadi di Ya'an, Tiongkok pada tahun 2013 sebagai studi kasus dan memperkenalkan metode pembelajaran mendalam untuk mengekstrak informasi emosional masyarakat yang terperinci dari data besar media sosial Tiongkok untuk membantu dalam analisis bencana. Dengan menggabungkan data ini dengan data informasi geografis lainnya (seperti data distribusi kepadatan penduduk, data POI (tempat tujuan), dll.), kami dapat membantu lebih lanjut dalam penilaian populasi yang terkena dampak, mengeksplorasi hukum pergerakan emosional, dan mengoptimalkan strategi mitigasi bencana.

7.1. PENDAHULUAN

Dengan popularitas perangkat seluler dan perkembangan infrastruktur jaringan, media sosial dengan cepat terintegrasi ke dalam kehidupan masyarakat. Orang-orang dapat dengan mudah membagikan apa yang mereka lihat dan dengar, bahkan apa yang mereka rasakan dan pikirkan melalui media sosial. Mereka seperti "sensor seluler" yang mengumpulkan informasi di sekitar mereka secara terus-menerus. Hal ini memberikan cara baru untuk memperoleh data terkait bencana. Dibandingkan dengan metode pengumpulan informasi bencana secara tradisional, media sosial memiliki karakteristik penyediaan informasi yang real-time dan berbiaya rendah. Selain itu, data tersebut banyak mengandung informasi geografis (seperti lokasi, waktu, dan informasi atribut lainnya) yang sangat penting untuk mitigasi bencana. Oleh karena itu, banyak peneliti yang menyadari pentingnya media sosial dalam mitigasi bencana. Mereka telah mempelajari bencana dari perspektif ekstraksi peristiwa, aturan lintasan pengguna dan fusi data, dll., dan mencapai hasil yang baik. Namun, hanya sedikit peneliti yang

menganggap informasi emosional publik yang terkandung dalam media sosial (terutama emosi yang sangat mendalam) sebagai atribut informasi geografis terkait bencana untuk membantu mitigasi bencana. Ketika bencana terjadi, emosi masyarakat sering kali mengungkapkan sikap masyarakat terhadap bencana, kebutuhan saat bencana, dan umpan balik mengenai bantuan bencana, dll. Hal ini sangat membantu untuk memahami perkembangan bencana dengan cepat dan meningkatkan efisiensi penyelamatan secara efektif. Namun, masih belum ada kerangka kerja yang efektif untuk mengumpulkan, memproses, dan menggunakan informasi emosional ini dengan cepat.

Ada tiga permasalahan yang terlibat: (1) Bagaimana kategori emosi masyarakat yang terperinci dapat dibagi selama bencana? (2) Media sosial memiliki basis pengguna yang besar. Kita ambil contoh mikro-blog Sina, sebuah media sosial asal Tiongkok. Menurut statistik, pada Q3 2018, platform media sosial mikro-blog Sina di Tiongkok memiliki lebih dari 431 juta pengguna aktif bulanan. Ketika terjadi bencana, hal ini akan menghasilkan banyak data terkait bencana. Dengan demikian, bagaimana informasi emosional terperinci yang terkandung dalam data ini dapat diekstraksi dengan lebih akurat? (3) Ketika emosi yang mendalam ini diekstraksi, bagaimana emosi tersebut dapat dianggap sebagai atribut informasi geografis terkait bencana untuk membantu mitigasi bencana? Dalam bab ini, kami menggunakan mikro-blog Sina dan mengambil bencana gempa bumi sebagai contoh untuk menggambarkan bagaimana kerangka kerja yang kami bangun mengekstraksi emosi masyarakat secara mendalam dan menggunakannya untuk melakukan mitigasi bencana.

Tidak seperti kebanyakan studi analisis emosi (mereka biasanya membagi emosi menjadi tiga kategori: positif, netral, dan negatif), kami membagi emosi masyarakat selama bencana ke dalam lebih banyak dimensi, karena penggunaan berbagai dimensi emosi dalam konteks bencana dapat memberikan gambaran yang lebih rinci. bencana yang akan dijelaskan. Selain itu, penelitian telah menggambarkan pentingnya informasi emosional multidimensi dalam bencana. Ekman, dkk menunjukkan perbedaan antara kemarahan, rasa jijik, ketakutan, dan kesedihan dalam hal peristiwa yang mendahuluinya dan kemungkinan respons perilaku. Oliver Gruebner dkk menganalisis bagaimana menerapkan berbagai dimensi emosi negatif (termasuk kemarahan, ketakutan, kesedihan, keterkejutan, kebingungan, rasa jijik) untuk mensurvei kesehatan mental bencana. Studi psikologis yang ada juga menyebutkan pembagian emosi yang sangat rinci dalam suatu bencana. Oleh karena itu, berdasarkan penelitian sebelumnya dan korpus yang digunakan dalam bab ini, kami membagi emosi negatif menjadi kemarahan, kecemasan, ketakutan, dan kesedihan.

Metode yang umum digunakan untuk klasifikasi emosi mencakup algoritma berbasis aturan dan model pembelajaran mesin tradisional. Algoritme berbasis aturan terutama menggunakan leksikon emosional tertentu dan aturan tata bahasa yang sesuai untuk menghitung intensitas emosional teks. Metode ini bergantung pada sejumlah besar operasi manual, seperti pengembangan aturan pencarian secara manual dan leksikon emosional skala besar, yang menentukan keakuratan metode. Selain itu, metode ini lemah dalam menangani kata-kata berhenti dan kata-kata baru. Sulit juga untuk menambahkan beberapa kata slang dan kata kunci Internet ke dalam leksikon emosional pada waktunya, seperti “喜大普奔”

(kepuasan besar), “狂顶” (sangat mendukung), dll., yang sering muncul di media sosial. Model pembelajaran mesin tradisional, seperti naif Bayes, entropi maksimum, dan mesin vektor dukungan tidak bergantung pada leksikon emosional atau aturan pencarian. Mereka hanya perlu memberi anotasi secara manual pada set pelatihan. Namun, metode pembelajaran mesin tradisional didasarkan pada model bag-of-words, yang mengabaikan hubungan semantik dalam teks. Dengan kata lain, tidak mempertimbangkan urutan kata dalam sebuah kalimat, sehingga dapat dengan mudah menyebabkan kesalahan klasifikasi emosi. Misalnya kalimat “Meskipun gempanya dahsyat, kami aman dan sehat” dan “Meskipun kami aman dan sehat, gempanya dahsyat” mengandung kata-kata yang sama, namun mengungkapkan emosi yang berbeda. Selain itu, untuk model pembelajaran mesin tradisional, masukannya adalah kata-kata fitur yang diambil dari teks setelah segmentasi. Definisi kata fitur memiliki dampak yang signifikan terhadap efisiensi model. Kami memilih metode pembelajaran mendalam untuk mengekstrak emosi publik dari media sosial. Dibandingkan dengan metode berbasis aturan, pembelajaran mendalam tidak bergantung pada leksikon emosional apa pun. Oleh karena itu, tidak terpengaruh oleh kata-kata baru dan tidak dikenal. Berbeda dengan pembelajaran mesin tradisional, pembelajaran mendalam menggunakan model vektor kata untuk menggantikan model kumpulan kata, yang dapat memanfaatkan informasi semantik dalam kalimat dengan baik. Banyak penelitian menunjukkan bahwa kinerja pembelajaran mendalam dalam tugas pemrosesan bahasa alami (NLP) lebih baik daripada pembelajaran mesin tradisional.

Selain itu, kami menggunakan emosi publik yang diekstraksi secara terperinci dan menggabungkannya dengan data informasi geografis tradisional (data distribusi kepadatan penduduk, data tempat tujuan (POI), dll.), dan fungsi analisis spasial yang kuat dari GIS (sistem informasi geografis). untuk membantu penanggulangan bencana. Menggabungkan informasi emosional masyarakat dapat menghasilkan manfaat berikut: (1) Dapat meningkatkan akurasi dan efisiensi penilaian bencana. Misalnya, dengan bantuan fungsi analisis spasial yang canggih dari GIS, data informasi geografis tradisional (seperti distribusi populasi, distribusi lalu lintas, dll.), dan data distribusi emosional dapat digabungkan untuk menilai populasi yang terkena dampak secara real-time. Orang yang mengungkapkan emosi negatif umumnya dianggap lebih mudah terkena dampak bencana. (2) Dapat membantu mengurangi kerugian akibat bencana. Misalnya, bencana, terutama bencana yang terjadi secara tiba-tiba (seperti gempa bumi, letusan gunung berapi, dll.), dapat dengan mudah menyebabkan masalah kesehatan mental terkait bencana, seperti gangguan stres pascatrauma (PTSD) dan depresi. Pemantauan tradisional mengalami kesulitan memperoleh informasi mengenai emosi masyarakat di wilayah bencana (walaupun sudah ada kuesioner, kinerja real-time-nya buruk). Jika informasi mengenai emosi masyarakat dan distribusi spatio-temporal yang terkait diketahui, departemen pengurangan bencana dapat mengambil tindakan penyelamatan psikologis yang sesuai untuk mengurangi terjadinya masalah kesehatan mental terkait bencana. Selain itu, bencana ekstrem mempunyai karakteristik yang tidak dapat dihindari dan tidak dapat diprediksi. Orang akan mengekspresikan emosi yang berbeda pada tahap yang berbeda dan memiliki respons berbeda untuk mencoba mengatasinya. Misalnya, orang yang cemas lebih

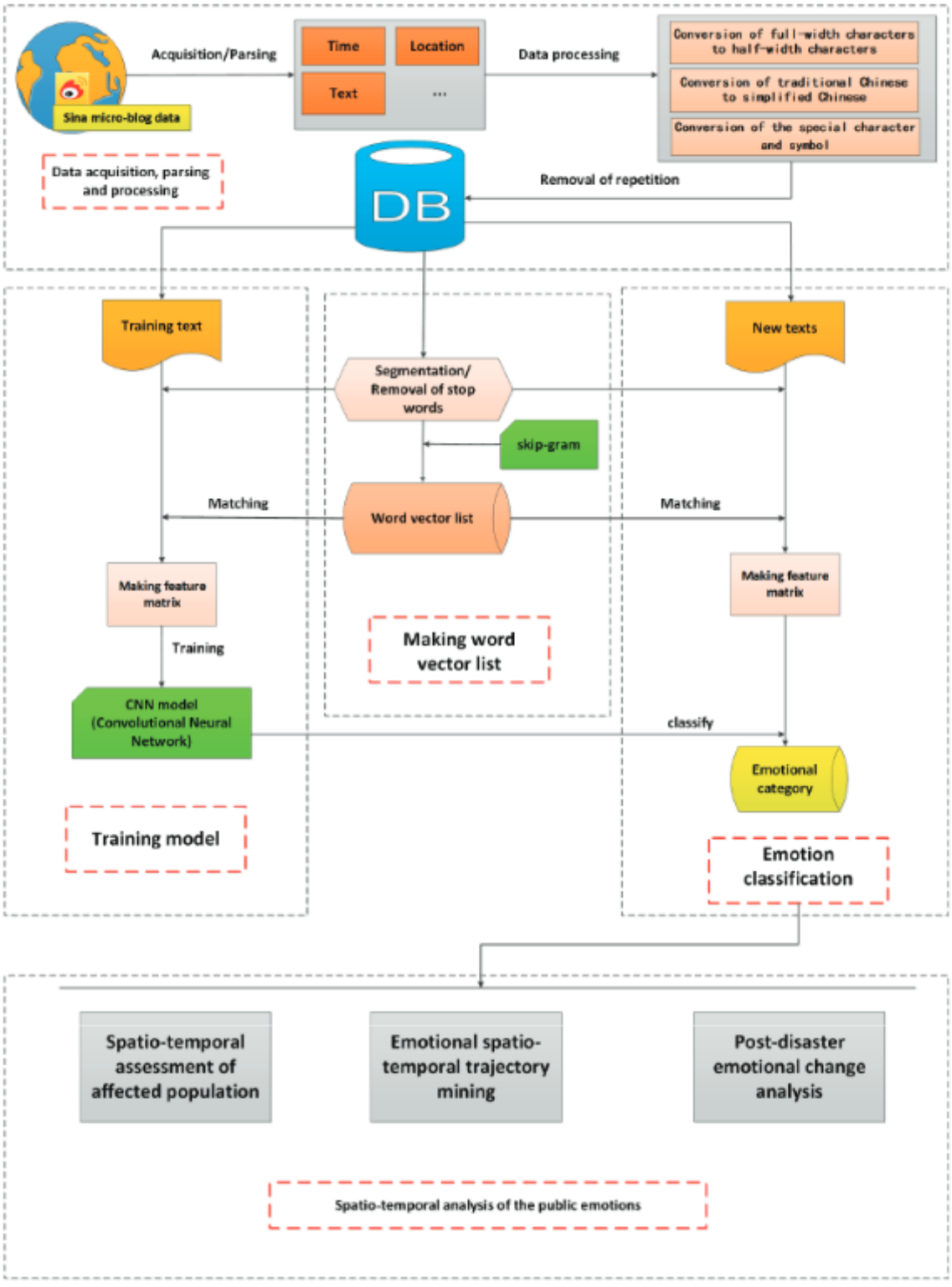
sensitif terhadap sisi negatif informasi terkait peristiwa dan mudah terpengaruh oleh rumor. Oleh karena itu, melalui pemahaman sebaran masyarakat yang cemas, kita dapat merilis informasi bencana yang benar pada waktu yang tepat untuk mencegah rumor yang mengganggu masyarakat yang cemas. (3) Mempelajari lebih banyak tentang penyebab emosi dapat membantu kita mengoptimalkan keputusan darurat. Dengan menggunakan kategori emosi yang berbeda, kita dapat mengeksplorasi penyebab emosi yang berbeda, seperti mengapa emosi marah lebih dominan di area tertentu dan emosi cemas lebih dominan di area lain, dan mengapa kategori emosi berubah di beberapa tempat seiring berjalannya waktu. Dengan memahami penyebab emosi, departemen pengurangan bencana dapat melakukan tindakan penanggulangan yang ditargetkan. Dalam analisis spatiotemporal informasi emosi masyarakat, kerangka kerja dalam bab ini mencakup penilaian populasi yang terkena dampak secara real-time, mengeksplorasi hukum pergerakan emosi, dan memantau penyebab perubahan emosi.

7.2. KERANGKA ANALISIS EMOSI MASYARAKAT DARI BIG DATA MEDIA SOSIAL

Kerangka kerja untuk menganalisis peran emosi masyarakat dalam mitigasi bencana yang diusulkan dalam bab ini mencakup lima fase utama: perolehan dan pemrosesan data, pembuatan daftar vektor kata, pelatihan model, klasifikasi emosi, dan analisis spatio-temporal emosi masyarakat (sebagai ditunjukkan pada Gambar 7.1).

Akuisisi dan Parsing Data Media Sosial

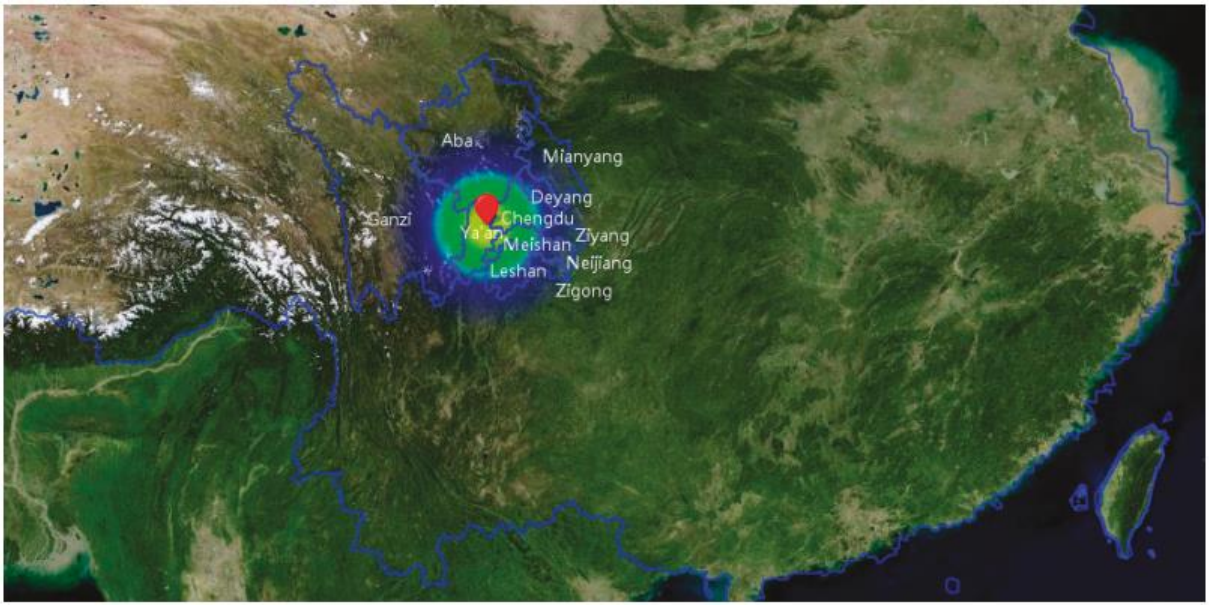
Kami menggunakan gempa bumi yang terjadi di Ya'an, Sichuan, Tiongkok, pada pukul 08:02 tanggal 20 April 2013, sebagai studi kasus. Menurut laporan China Seismograph Network (http://news.ceic.ac.cn/CC20130420080246.html), kekuatan gempa ini berkekuatan 7,0 dan kedalaman fokusnya 13 kilometer. Episentrum gempa ini terletak pada 30.30◦ LU, 103.00◦ BT yang menyebabkan sekitar 1,52 juta orang akan terkena dampaknya di wilayah seluas 12.500 kilometer persegi.



Gambar 7.1. Kerangka klasifikasi emosi otomatis dan analisis bencana.

Dalam buku ini, data media sosial diperoleh dari mikroblog Sina dari wilayah sekitar pusat gempa dengan radius 200 km yang rusak parah akibat gempa. Kota-kota yang terkena dampak termasuk Ya'an, Meishan, Ganzi, Leshan, Ziyang, Deyang, Chengdu, Aba, Zigong, Mianyang, dan Neijiang, seperti yang ditunjukkan pada Gambar 7.2. Rentang waktu data

media sosial adalah dari 20 April hingga 26 April. 2017. Platform media sosial biasanya menyediakan antarmuka atau API (Application Programming Interface) yang memungkinkan pengembang mengambil data media sosial. Namun, pengambilan data dengan cara ini memiliki keterbatasan yang besar; misalnya, Anda tidak dapat mengatur rentang waktu dan topik, dll. Oleh karena itu, dalam bab ini, kami menggunakan kemampuan pencarian lanjutan mikro-blog Sina untuk mendapatkan data dengan menggunakan rentang waktu, nama kota, dan kata kunci terkait peristiwa.



Gambar 7.2. Wilayah studi gempa Ya'an tahun 2013

Format data awalnya adalah bahasa markup hypertext (HTML). Kami menguraikan data ke dalam format data terstruktur termasuk kolom seperti "waktu", "lokasi", "teks", dll. Diantaranya, lokasi diwakili oleh alamat dan keakuratannya berbeda-beda. Kami mengambil Chengdu sebagai contoh. Beberapa alamat dijelaskan secara lebih rinci, seperti "Gerbang Timur Universitas Sichuan", "Jalan Sishengci Utara", dll. Beberapa alamat dijelaskan secara kasar, seperti "Distrik Baru Funan". Ada juga beberapa teks yang tidak memiliki informasi alamat. Alasannya adalah kebiasaan penggunaan orang berbeda-beda (beberapa orang tidak mau membagikan informasi lokasinya). Kami menggunakan API yang disediakan oleh Baidu (http://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-geocoding) untuk mengonversi alamat ini menjadi lintang dan bujur. Diantaranya, untuk data garis, seperti "Jalan Sishengci Utara", kami mengambil koordinat titik tengahnya untuk merepresentasikannya. Untuk data permukaan, seperti "Kampus Wangjiang Universitas Sichuan" dan bahkan "Distrik Baru Funan", kami mengekstrak koordinat titik pusat untuk mewakili masing-masing data tersebut. Kami tidak menetapkan koordinat pada teks-teks yang tidak memiliki informasi alamat, termasuk teks-teks dengan alamat kasar. Mereka hanya diberi label "Chengdu."

7.3 PEMROSESAN DATA MEDIA SOSIAL

Pada langkah pemrosesan selanjutnya, kami terutama menangani data teks. Langkah-langkah pemrosesan teks utama mencakup konversi karakter lebar penuh menjadi karakter lebar setengah dan dari bahasa Mandarin tradisional ke bahasa Mandarin sederhana, serta pengenalan karakter dan simbol khusus. Tujuan dari dua langkah pertama adalah untuk meningkatkan efisiensi komputasi model. Langkah ketiga bertujuan untuk mengenali karakter dan simbol khusus, seperti "(>_<)", "Ƃ", yang dihapus dan diabaikan oleh banyak alat pemrosesan bahasa alami (NLP) yang umum. Namun untuk analisis emosional, karakter dan simbol khusus tersebut memiliki makna emosional, misalnya "(>_<)" dan "(>_<)>" dapat mengungkapkan emosi yang bermasalah. Oleh karena itu, dalam tulisan ini, kami menafsirkannya menjadi teks yang dapat diproses oleh NLP. Beberapa karakter dan simbol khusus dapat diterjemahkan ke dalam teks oleh platform mikro-blog. Misalnya, 😭 bisa diterjemahkan menjadi "air mata" (泪). Namun, emotikon lain yang tidak dapat diuraikan oleh platform mikro-blog, seperti (>_<) dan 💔 , diinterpretasikan berdasarkan "daftar emotikon" di web, yang mencakup implikasi emosional dari semua jenis emotikon dalam jumlah besar. dari literatur yang diterbitkan. Misalnya, (>_<) dapat diterjemahkan menjadi "bermasalah" (焦虑) dan 💔 dapat diterjemahkan menjadi "sedih" (伤心). Terakhir, setelah menghilangkan duplikasi, ada 39341 catatan data yang tersimpan di database kami.

Membangun Daftar Vektor Kata

Pertama-tama kami mengubah setiap kata dari teks yang diproses sebelumnya menjadi vektor multidimensi. Proses ini mencakup dua fase: segmentasi kata dan penghapusan kata-kata berhenti, dan pembuatan daftar vektor kata.

Segmentasi Kata dan Penghapusan Stop Words

Berbeda dengan bahasa Inggris, tidak ada pemisah spasi di antara kata-kata berbahasa Mandarin. Oleh karena itu, kami perlu mengelompokkan teks berbahasa Mandarin untuk mendapatkan kata-kata yang terpisah. Selain itu, mikro-blog Tiongkok lebih bersifat sehari-hari, yang membawa tantangan besar pada segmentasi kata. Kami membandingkan banyak alat segmentasi kata berbahasa Mandarin yang berbeda, seperti "Stanford NLP", "ANSJ", "NLPIR (Natural Language Processing & Information Retrieval Sharing Platform)", dan seterusnya. Kami menemukan bahwa "NLPIR" memiliki kinerja terbaik dalam hal akurasi dan kecepatan segmentasi kata.

Ada banyak kata yang tidak berarti dalam teks setelah segmentasi kata; ini disebut kata-kata berhenti, seperti "在 (on)," "是 (is)," "一会 (a moment)," dan seterusnya. Kata-kata ini dapat mempengaruhi keakuratan model dan oleh karena itu harus dihilangkan. Dalam bab ini, kami menggunakan kosakata kata-kata berhenti yang dikembangkan oleh Institut Teknologi Harbin—Pusat Penelitian Komputasi Sosial dan Pengambilan Informasi untuk menghilangkan kata-kata berhenti. Karena fokus bab ini adalah pada analisis emosional, kami mengoptimalkan kosakata kata-kata penghenti dengan menghilangkan kata-kata sentimental, seperti "愤然 (marah)," "幸亏 (untungnya)," "嘻 (hei)", dll.

Konstruksi Daftar Vektor Kata

Input yang digunakan untuk model klasifikasi emosi adalah matriks vektor kata. Kami perlu mengubah setiap kata dalam teks mikro-blog menjadi vektor multidimensi, dan kemudian mengubah seluruh kalimat menjadi matriks vektor kata. Dalam bab ini, kami mengubah semua teks yang diproses sebelumnya menjadi daftar vektor kata. Teks pelatihan dan teks baru yang akan dikategorikan diubah menjadi matriks vektor kata dengan cara mencocokkannya dengan daftar vektor kata. Metode yang kami gunakan untuk ini adalah word2vec, yang memproyeksikan setiap kata dalam setiap kalimat ke ruang vektor berdimensi tertentu.

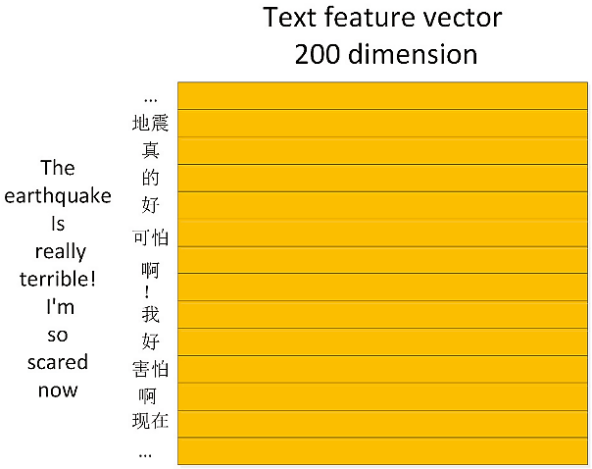
Ada dua model yang umum digunakan di word2vec, yaitu skip-gram dan CBOW (Continuous Bag-of-Words). Sejumlah besar percobaan telah dilakukan untuk membandingkan kedua model ini dalam hal kinerja dan akurasi dan hasilnya menunjukkan bahwa tingkat akurasi semantik model skip-gram lebih baik dibandingkan model CBOW. Oleh karena itu, kami menggunakan model Skip-gram dalam percobaan kami untuk membuat vektor fitur teks.

Model skip-gram dapat menentukan korelasi antar kata untuk pelatihan korpus. Korelasi ini diwakili oleh vektor fitur multidimensi dari setiap kata. Selain itu, vektor fitur multidimensi ini dihitung dengan mempertimbangkan sepenuhnya konteks informasi semantik. Dari rumus di bawah ini, jika diberi kata saat ini, wi, model ini mencoba menemukan kata yang memiliki hubungan semantik kontekstual dengan kata saat ini. Sasaran model ini adalah memaksimalkan fungsi tujuan, G: (Persamaan 1)

$$G = \sum_{w_i \epsilon C} \log P\left(Context(w_i)|w_i\right)$$

Dalam rumus ini, wi mewakili kata saat ini dan C mewakili jendela konteks. $P(Context(w_i)|w_i)$ mewakili probabilitas informasi konteks dalam kata saat ini.

Ketika pelatihan menyatu, kata-kata dengan makna semantik serupa berada lebih dekat dalam ruang vektor dimensi yang ditentukan. Kami mengekspor vektor fitur teks dari setiap kata dalam korpus pelatihan untuk menghasilkan penyematan kata. Struktur vektor fitur teks ditunjukkan seperti Gambar 7.3.



Gambar 7.3. Struktur vektor fitur teks.

Pelatihan Model

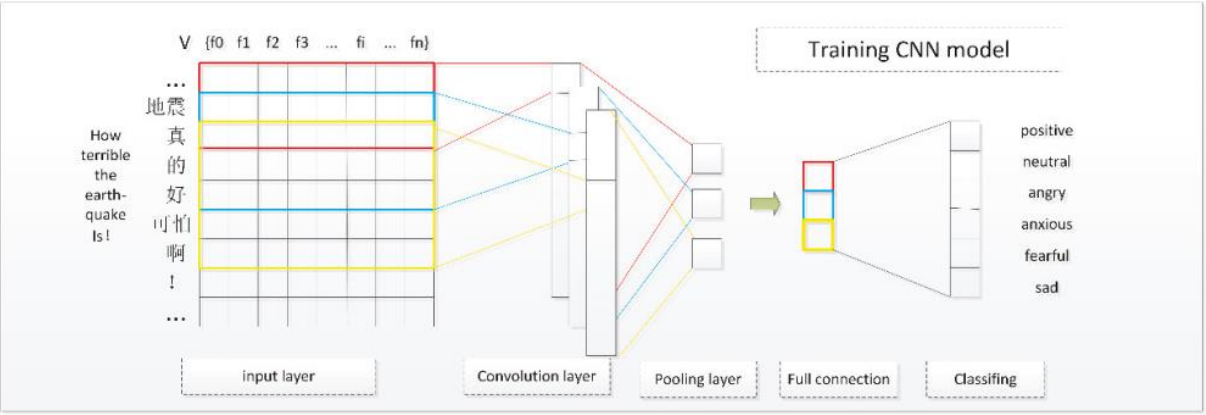
Model deep leaning yang dipilih dalam bab ini adalah jaringan saraf konvolusional. Kami membaca banyak literatur terkait dan menemukan bahwa metode pembelajaran mendalam yang berbeda dapat dipilih untuk klasifikasi emosi, seperti jaringan saraf konvolusional (CNN), jaringan saraf berulang (RNN), jaringan perhatian hierarki (HAN), dll. Model-model ini semuanya memiliki karakteristik dan skenario penggunaannya masing-masing. Menurut literatur, CNN melakukan klasifikasi emosi dengan baik, terutama dalam kalimat yang lebih pendek. RNN melakukan klasifikasi emosi tingkat dokumen dengan baik. Penelitian sebelumnya menyajikan kinerja CNN, RNN, dan HAN dalam klasifikasi emosi. Hasil penelitian menunjukkan bahwa ketika korpus pelatihan cukup besar, HAN memiliki akurasi tertinggi, namun CNN memberikan kinerja terbaik ketika korpus pelatihan tidak terlalu besar. Anotasi korpus pelatihan yang besar membutuhkan banyak tenaga dan waktu. Selain itu, pelatihan model HAN dan RNN membutuhkan waktu lebih lama dibandingkan model CNN. Dalam bab ini, korpus pelatihan yang digunakan adalah teks mikro-blog, yang sebagian besar berupa teks pendek. Selain itu, jumlah data korpus pelatihan yang diberi tag secara manual lebih sesuai untuk model CNN. Oleh karena itu, kami memilih CNN sebagai metode untuk mengekstrak emosi masyarakat yang terdapat di media sosial. Proses pelatihan model ditunjukkan di bawah ini.

Segmentasi Kata, Penghapusan Stop Word, dan Konstruksi Matriks Fitur

Pertama, kami mengelompokkan teks pelatihan untuk mendapatkan kata-kata terpisah. Kemudian, kami menggunakan kosakata kata-kata berhenti untuk menghilangkan kata-kata tidak bermakna yang terkandung dalam kata-kata terpisah tersebut. Terakhir, sisa kata diubah menjadi vektor kata melalui pencocokan dengan daftar vektor kata yang dibuat sebelumnya. Pada akhirnya, setiap kalimat diubah menjadi matriks fitur.

7.4 PELATIHAN MODEL JARINGAN NEURAL KONVOLUSIONAL

Jaringan saraf konvolusional (CNN) adalah varian dari jaringan saraf. Ini pertama kali berhasil digunakan untuk pengenalan gambar dan video. Belakangan, beberapa peneliti memperkenalkannya ke dalam bidang pemrosesan bahasa alami dan menemukan bahwa hal ini mempunyai efek yang baik. Model CNN yang digunakan dalam bab ini terdiri dari lapisan masukan, lapisan konvolusional, lapisan penyatuan, lapisan terhubung penuh, dan klasifikasi. Struktur CNN ditunjukkan pada Gambar 7.4.



Gambar 7.4. Struktur model jaringan saraf konvolusional (CNN)

Selama proses pelatihan model CNN, neuron di dalamnya biasanya diatur ke tiga dimensi: kedalaman, lebar, dan tinggi. Ukuran masing-masing lapisan adalah kedalaman × lebar × tinggi. Misalnya, jika ada kalimat dengan 140 kata, dan setiap kata diatur menjadi 200 dimensi, maka ukuran lapisan masukan adalah 1 × 140 × 200. Selanjutnya, kami memperkenalkan lapisan jaringan saraf konvolusi.

Lapisan masukan: Lapisan masukan CNN adalah matriks yang terdiri dari vektor fitur teks. Matriks ini dihitung menggunakan model skip-gram. Baris dan kolom (dimensi) dalam matriks ini telah diatur sebelum kita memasukkan matriks tersebut ke dalam model jaringan saraf. Mengambil contoh teks mikro-blog Sina, jumlah karakter dalam setiap kalimat kurang dari 140. Oleh karena itu, kita mengatur baris dalam matriks menjadi 140. Jika jumlah kata dalam sebuah kalimat kurang dari 140 karakter, kita menggunakan karakter kosong "spasi" untuk melengkapi karakter yang hilang. Oleh karena itu, setiap kalimat diungkapkan sebagai berikut: (Persamaan 2):

$$S_{1:140} = S_1 \oplus S_2 \oplus S_3 \ldots \oplus S_{140}$$

Dalam rumus ini, S mewakili karakter atau “pad”, dan $\oplus$ adalah operator penggabungan. Lapisan konvolusional: Lapisan konvolusi terutama digunakan untuk mengekstrak fitur. Ini mengabstraksi beberapa elemen yang terfragmentasi menjadi fitur yang dapat digunakan untuk membedakan kategori yang berbeda. Melalui konvolusi, banyak fitur tingkat rendah dapat diabstraksikan ke fitur tingkat yang lebih tinggi. Misalnya, satu kata “打” atau “panggilan” tidak memiliki makna emosional. Namun, fitur tingkat yang lebih tinggi “打 panggilan (pujian)” dapat mengekspresikan atribut emosional. Atribut emosional dari kata-kata ini dapat diperoleh model melalui sejumlah besar korpus pelatihan.

Diberikan matriks u yang berasal dari lapisan masukan untuk operasi konvolusi, rumusnya adalah sebagai berikut: (Persamaan 3)

$$c_j = f\left(\boldsymbol{u} * \boldsymbol{k}_j + \boldsymbol{b}_j\right)$$

Untuk matriks $\mathbf{u} \in \mathbb{R}^{D \times L}$, D mewakili dimensi penyematan, dan L mewakili panjang kalimat. Parameter $\mathbf{k} \in \mathbb{R}^{D \times s}$ mewakili kernel konvolusional ke-j, yang diterapkan pada jendela kata-

kata s. Parameter $b_j \in \mathbb{R}$ mewakili suku bias. $f(\boldsymbol{u} * \boldsymbol{k}_j + \boldsymbol{b}_j)$merupakan fungsi aktivasi non-linier.

Lapisan pengumpulan: Setelah operasi konvolusi, kita dapat menggunakan fitur keluaran untuk mengklasifikasikan emosi secara langsung. Namun, dalam melakukan hal ini, kita tidak hanya akan menghadapi tantangan kompleksitas komputasi, namun juga masalah over-fitting, yang akan mempengaruhi keakuratan klasifikasi. Operasi pooling dapat mengatasi permasalahan tersebut dengan baik. Selain itu, operasi pengumpulan juga dapat berfungsi sebagai pemilih fitur yang dapat membantu mengidentifikasi fitur-fitur terpenting untuk meningkatkan kinerja klasifikasi.

Ada dua metode yang bisa dipilih yaitu max pooling dan average pooling. Kami mencapai hasil yang lebih baik dengan metode pengumpulan maksimal. Metode ini memilih fitur semantik global dan berupaya menangkap fitur terpenting dengan nilai tertinggi untuk setiap peta fitur [37]. Output dari operasi konvolusi c_j digunakan sebagai input operasi pooling. Rumusnya adalah sebagai berikut: (Persamaan 4)

$$p_j = pooling(c_j) + b_j$$

Lapisan Terhubung Sepenuhnya: Neuron pada lapisan ini memiliki koneksi penuh dengan semua neuron pada lapisan sebelumnya. Sedangkan nilai full connection layer dapat dihitung melalui neuron-neuron pada layer sebelumnya. Dalam proses perhitungannya biasanya digunakan metode regularisasi dropout untuk menghindari over-fitting.

Klasifikasi: Kita bisa mendapatkan label emosional dari teks asli melalui fungsi softmax. Dengan kata lain, hasil perhitungan ini mewakili distribusi probabilitas dari label emosional. Berdasarkan korpus pelatihan, kita dapat mengidentifikasi parameter terbaik untuk model CNN. Kemudian, model terlatih ini dapat digunakan untuk menghitung kategori emosional teks baru.

7.5 KLASIFIKASI EMOSI

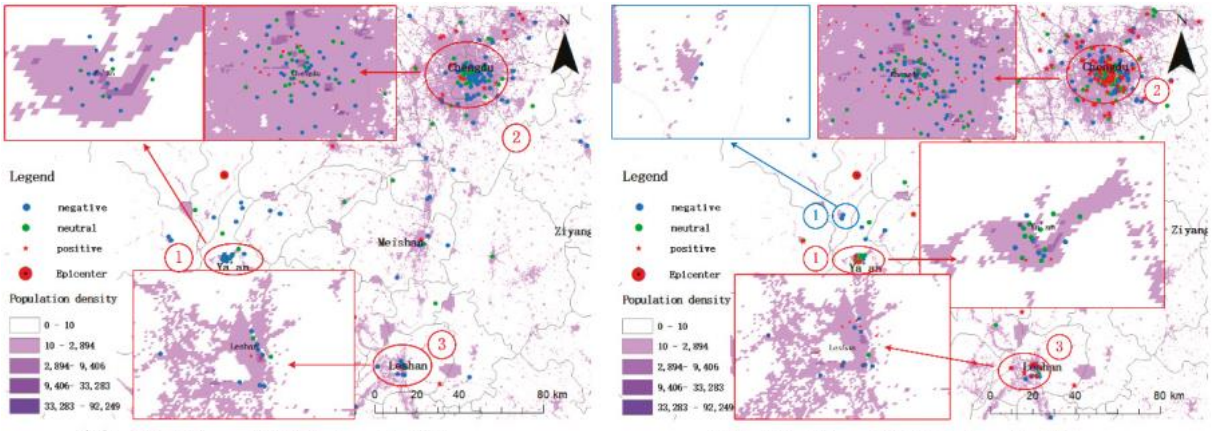
Kami menggunakan model CNN terlatih untuk menganalisis teks baru. Emosi yang terkandung dalam teks-teks tersebut dibagi menjadi enam kategori: positif, netral, marah, cemas, takut, dan sedih. Diantaranya, emosi positif tersebut terutama mencakup kepuasan masyarakat terhadap bantuan bencana, keinginan masyarakat terhadap lokasi bencana, dan kegembiraan karena bisa selamat. Emosi netral terutama mencakup deskripsi objektif mengenai bencana tersebut. Dalam proses klasifikasi, teks baru diolah terlebih dahulu dengan menggunakan segmentasi kata dan penghilangan stopword. Kemudian, daftar vektor kata yang telah dilatih sebelumnya digunakan untuk menerjemahkan setiap kata menjadi vektor kata. Selanjutnya setiap teks baru ditransformasikan ke dalam matriks vektor kata. Terakhir, matriks vektor kata dimasukkan ke dalam model CNN yang dilatih. Melalui penghitungan model, setiap teks baru diberi label ke dalam kategori emosional yang berbeda. Kami mengklasifikasikan seluruh 39.344 potongan teks ke dalam enam kategori emosi berdasarkan proses klasifikasi ini.

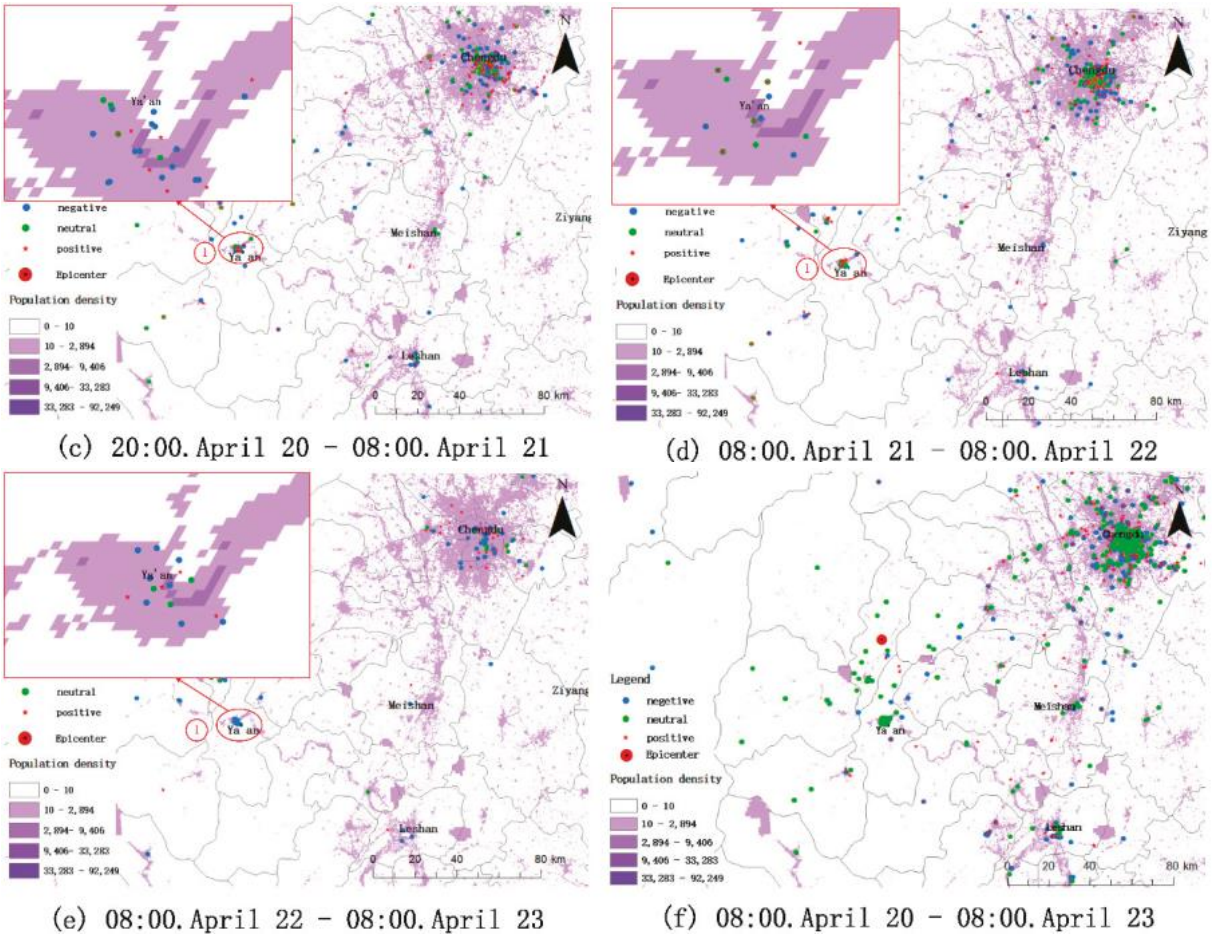
Analisis Spatio-Temporal Emosi Masyarakat

Kerangka dalam tulisan ini bertujuan untuk membantu mitigasi bencana dengan menggunakan informasi emosional masyarakat yang terdapat di media sosial. Dalam prosesnya, informasi emosional dianggap sebagai atribut informasi geografis. Kemampuan analisis spasial yang kuat dari GIS digunakan untuk menggabungkan informasi emosional dengan data geografis lainnya untuk menggali pengetahuan yang lebih berguna. Misalnya, data distribusi kepadatan penduduk dapat ditambahkan untuk melakukan penilaian spatio-temporal terhadap populasi yang terkena dampak. Data POI (seperti suaka) dapat dianggap mengeksplorasi hukum lintasan spatio-temporal masyarakat dalam bencana mendadak. Selain itu, informasi emosional juga dapat membantu departemen pengurangan bencana untuk menyaring kebutuhan masyarakat yang mendesak dari sejumlah besar informasi. Tuntutan masyarakat yang mengandung informasi emosional juga merupakan umpan balik yang efektif untuk upaya pengurangan bencana. Mereka dapat membantu mengoptimalkan pengambilan keputusan untuk meningkatkan efisiensi penyelamatan.

Analisis Spatio-Temporal Informasi Emosional Publik

Sangat penting untuk mengetahui sebaran penduduk yang terkena dampak pada saat gempa terjadi. Hal ini dapat membantu memastikan penilaian yang efektif terhadap situasi bencana dan penyebaran sumber daya penyelamatan yang rasional. Pada bagian ini, kami menggabungkan data sebaran kepadatan penduduk terkait wilayah penelitian dengan informasi spatio-temporal yang terdapat di media sosial untuk membantu analisis. Diantaranya, data sebaran kepadatan penduduk diambil dari GHSL (Global Human Settlement Layer). Pengenalan informasi emosional masyarakat dapat meningkatkan akurasi penilaian. Secara umum, emosi negatif diyakini menunjukkan bahwa gempa bumi mempunyai dampak yang lebih besar terhadap masyarakat. Selanjutnya, berdasarkan aturan pembagian periode waktu dari departemen penyelamatan, kami menetapkan enam periode waktu, yaitu: 0–4 jam, 4–12 jam, 12–24 jam, 24–48 jam, 48–72 jam, dan 0–72 jam setelah gempa. Kemudian, kami menggunakan analisis overlay perangkat lunak GIS untuk memproses data ini pada setiap periode waktu. Hasil analisis data terkait ditunjukkan pada Gambar 7.5.

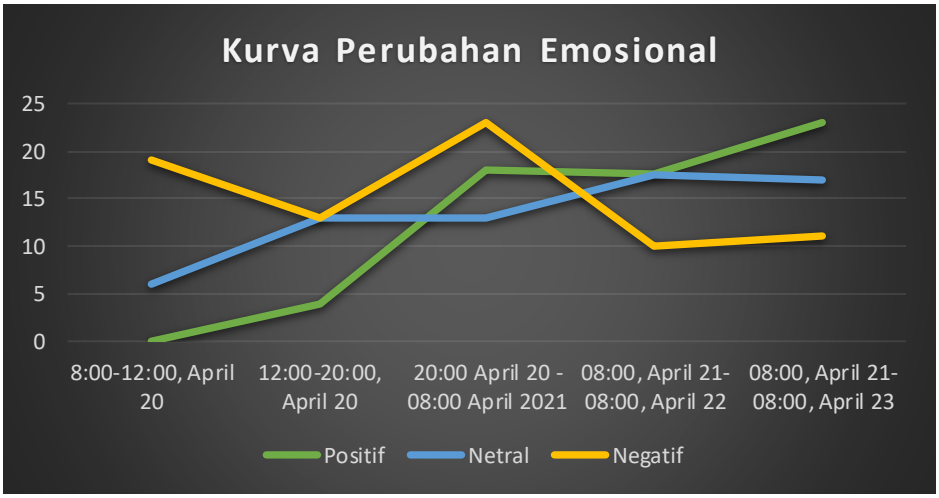




Gambar 7.5. Karakteristik distribusi emosional dari populasi yang terkena dampak. Gambar (a), (b), (c), (d) dan (e) menggambarkan distribusi emosi dalam periode waktu yang berbeda dalam waktu 72 jam setelah bencana. Gambar (f) menunjukkan distribusi emosi selama 72 jam. Diantaranya, masing-masing lingkaran merah 1, lingkaran merah 2, dan lingkaran merah 3 pada gambar mewakili luas yang sama. Lingkaran biru 1 pada (b) menunjukkan bahwa dibandingkan dengan (a), emosi negatif baru muncul di area yang sama.

Kita tahu bahwa: (1) Volume data mikroblog lebih besar di tempat-tempat dengan kepadatan penduduk yang tinggi setelah gempa bumi dan emosi negatif mendominasi. (2) Dalam waktu empat jam setelah gempa, seperti terlihat pada Gambar 7.5a, hampir tidak ada emosi positif di wilayah dekat pusat gempa. Daerah yang jauh dari pusat gempa, seperti Leshan (lingkaran merah 3) dan Chengdu (lingkaran merah 2), mempunyai emosi positif yang lebih sedikit. (3) Dari 4 jam hingga 12 jam setelah gempa, seperti terlihat pada Gambar 7.5b, dibandingkan dengan sebaran informasi emosional sebelumnya, muncul beberapa emosi negatif baru di dekat pusat gempa, seperti lingkaran biru 1. Hal ini menunjukkan bahwa seiring berjalannya waktu dan seterusnya, beberapa kerusakan akibat bencana baru mungkin terjadi di wilayah ini. Kami memeriksa teks terkait dan menemukan bahwa titik emosi baru ini sebagian besar adalah kecemasan. Alasan masyarakat mengungkapkan kegelisahannya adalah karena "Jalan Fan Min" terhalang oleh batu-batu besar, dan masyarakat khawatir kendaraan

penyelamat yang tidak mengetahui informasi tersebut akan tertunda akibat kejadian tersebut. Emosi di area lain pada gambar ini juga telah berubah. Misalnya, dibandingkan dengan lingkaran merah 2 dan lingkaran merah 3 pada Gambar 7.5a, emosi di area Gambar 7.5b ini meningkat secara signifikan. Hal ini menandakan bahwa perhatian masyarakat terhadap gempa bumi terus meningkat pada periode ini. (4) Kami memilih wilayah dengan kepadatan penduduk tinggi di dekat pusat gempa untuk menganalisis secara detail bagaimana emosi berubah seiring berjalannya waktu. Daerah yang dipilih terletak di Distrik Yucheng di Ya'an dan ditandai dengan lingkaran merah 1 pada Gambar 7.5a–e. Gambar 7.6 menunjukkan perubahan volume data kategori emosi di area ini untuk periode waktu yang berbeda. Kami menemukan bahwa emosi positif mulai muncul pada periode kedua dan kemudian terus meningkat. Pasalnya, seiring dengan berjalannya operasi penyelamatan, rasa terima kasih masyarakat terhadap tim penyelamat semakin meningkat. Jumlah emosi negatif meningkat paling banyak pada periode ketiga dan kemudian berkurang secara bertahap. Meski periode ini hanya berlangsung 12 jam, jumlah emosi negatif merupakan yang tertinggi dari semua periode. Karena periode ini adalah malam pertama setelah gempa bumi, sebagian besar masyarakat sangat membutuhkan bantuan, seperti tenda, pakaian, dan lain-lain. Oleh karena itu, rasa cemas sangat dominan. Jumlah emosi netral yang diungkapkan tidak banyak berubah. Mereka terutama menggambarkan perkembangan gempa bumi. (5) Gambar 7.5f menggambarkan situasi keseluruhan 72 jam setelah gempa bumi. Kami menemukan bahwa kepadatan penduduk di Ya'an tidak tinggi, dan persebaran penduduk tidak seragam. Namun jumlah emosi yang diungkapkan di kota ini cukup banyak dan sebarannya cukup seragam, terutama mengenai emosi negatif. Hal ini menunjukkan bahwa dampak gempa bumi di Ya'an paling parah. Selain itu, Chengdu merupakan ibu kota Provinsi Sichuan dan memiliki kepadatan penduduk terbesar. Saat gempa terjadi, banyak emosi negatif yang diungkapkan di kota ini. Meskipun distribusi emosi ini tidak seragam, perhatian lebih harus diberikan pada hal ini untuk menghindari kecelakaan yang tidak terduga, seperti orang yang tersakiti oleh rumor karena emosi cemas. Pendekatan yang sama juga dapat diterapkan di kota-kota lain yang terkena dampak.

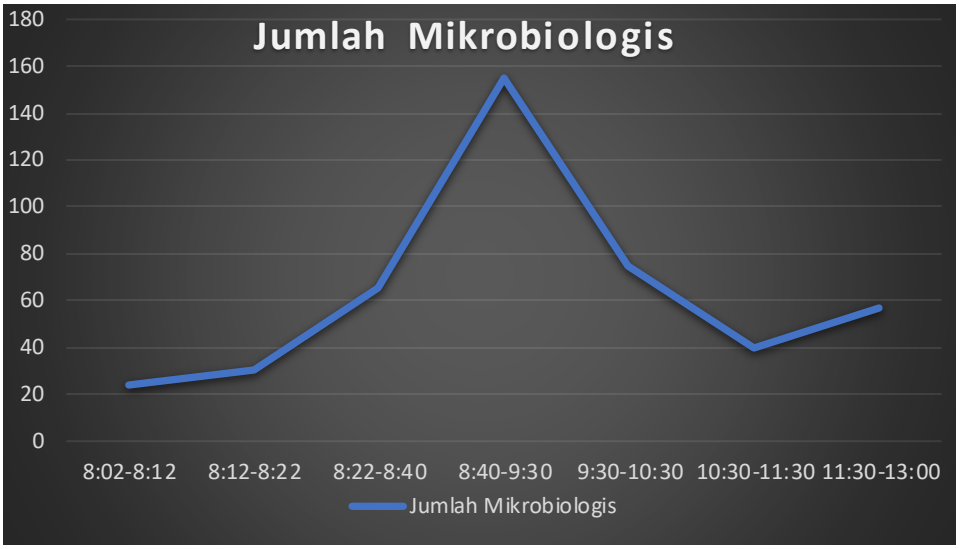


Gambar 7.6. Perubahan kategori emosi yang berbeda dalam volume data untuk periode waktu yang berbeda.

Pada bagian ini, kami melakukan analisis overlay yang berisi informasi emosional masyarakat dan distribusi kepadatan penduduk untuk menilai populasi yang terkena dampak. Karakteristik distribusi spatio-temporal informasi emosional publik dapat meningkatkan akurasi penilaian dan memberi kita informasi yang lebih berharga. Meskipun informasi emosional ini tidak tersebar secara merata, dan bahkan beberapa daerah dengan kepadatan penduduk yang tinggi hanya memiliki sedikit emosi negatif, seperti daerah yang berada di lingkaran biru 1 pada Gambar 7.5b, kita tetap harus memperhatikannya karena emosi yang diungkapkan oleh pengguna Penggunaan media sosial juga dapat mencerminkan emosi dari tetangga atau komunitas di sekitarnya, meskipun tetangga atau komunitas tersebut tidak menggunakan media sosial.

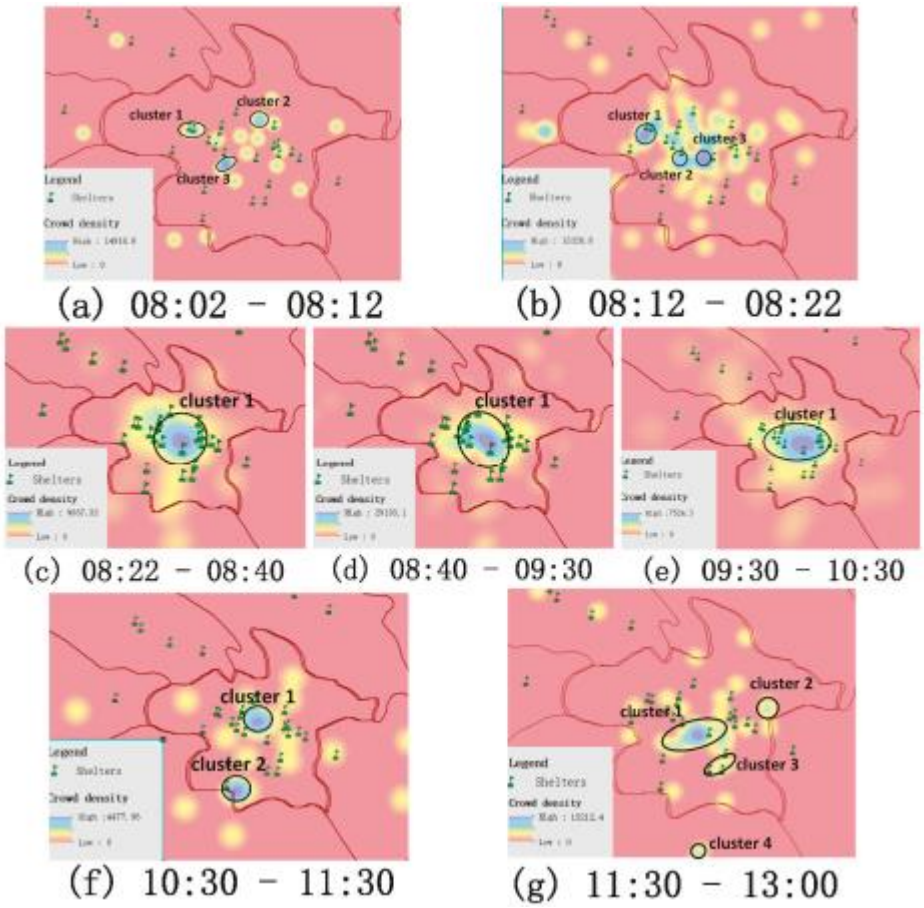
7.6 PENAMBANGAN LINTASAN SPATIO-TEMPORAL EMOSIONAL

Gempa bumi, sebagai bencana yang terjadi secara tiba-tiba, menimbulkan kerusakan yang sangat besar dalam jangka waktu yang singkat, dan sangat berbahaya bagi kehidupan manusia. Oleh karena itu, di sebagian besar kota terdapat banyak tempat berlindung bagi masyarakat untuk menghindari bencana tersebut. Pada bagian ini, kita mengeksplorasi bagaimana lintasan spatio-temporal manusia berubah ketika terjadi bencana mendadak dan apakah perubahan ini berkaitan dengan lokasi perlindungan. Lebih jauh lagi, dalam proses perubahan ini, kami menyelidiki kategori emosi mana yang ditunjukkan oleh manusia dan bagaimana kategori emosi tersebut berubah. Kami menggunakan Chengdu sebagai contoh dan menentukan lokasi penampungan di kota ini dari “Situs Web Resmi Pemerintahan Rakyat Kota Chengdu (http://cdtf.gov.cn/chengdu/smfw/csyjbn.shtml).” Kemudian, kami menerjemahkan shelter ini ke dalam koordinat melalui API Baidu (http://api.map.baidu.com/lbsapi/getpoint/index.html) dan membuat vektor peta wilayah ini. Mengingat terjadinya gempa bumi secara tiba-tiba, kami menetapkan tujuh periode waktu kecil, yaitu pukul 08:02 hingga 08:12, pukul 08:12 hingga 08:22, 08:22 hingga 08:40, 08:40 pukul 09:30, 09:30 hingga 10:30, 10:30 hingga 11:30, dan 11:30 hingga 13:00 (gempa terjadi pada pukul 08:02), untuk menganalisis perubahan kelompok dalam periode waktu yang terperinci ini. Gambar 7.7 menunjukkan perubahan jumlah orang dari waktu ke waktu. Kita dapat melihat bahwa pertumbuhan populasi paling cepat terjadi pada pukul 08:40 hingga 09:30. Hal ini mungkin mencerminkan bahwa sejumlah besar orang telah mencapai tempat penampungan terdekat selama periode ini.



Gambar 7.7. Perubahan jumlah kerumunan pada setiap periode waktu kecil.

Kami menggunakan algoritma kepadatan kernel untuk memvalidasi perubahan agregasi kerumunan dari waktu ke waktu dan mengeksplorasi hubungan antara pusat kepadatan kerumunan dan lokasi tempat penampungan, seperti yang ditunjukkan pada Gambar 7.8. Pada Gambar 7.8a, tiga cluster terbentuk dalam 10 menit pertama setelah gempa dan kami menandainya masing-masing sebagai cluster 1, cluster 2, dan cluster 3. Meski kepadatan cluster 1 kecil, namun terlihat intinya berada di lokasi shelter. Hal ini menunjukkan bahwa dalam waktu 10 menit setelah gempa, sejumlah kecil orang telah berkumpul di tempat penampungan terdekat. Cluster 2 dan cluster 3, khususnya cluster 3, memiliki kepadatan yang lebih tinggi dibandingkan cluster 1. Namun masyarakat di wilayah tersebut belum berkumpul di shelter. Sepuluh menit kemudian, antara pukul 08:12 hingga 08:22, jumlah orang bertambah seiring mereka semakin berkumpul di tempat penampungan. Kami menemukan bahwa beberapa shelter telah mengumpulkan banyak orang, seperti cluster 1 dan cluster 3, seperti yang ditunjukkan pada Gambar 8b. Seiring berjalannya waktu, sejumlah besar cluster kecil dan terpisah digabungkan menjadi cluster yang lebih besar, seperti yang ditunjukkan pada Gambar 7.8c – e. Hal ini menunjukkan bahwa selama periode tersebut, sejumlah besar orang telah mencapai tempat penampungan. Di antara mereka, jumlah orang antara pukul 08:40 hingga 09:30 adalah yang terbesar, seperti yang ditunjukkan pada Gambar 7.8d. Kemudian, jumlah orang yang berkumpul di tempat penampungan mulai berkurang. Namun massa belum bubar saat ini, seperti terlihat pada Gambar 7.8e. Kita dapat memahami perubahan ini melalui nilai kepadatan kerumunan yang sesuai. Pada Gambar 7.8f,g terlihat bahwa cluster besar mulai terpecah dan terurai menjadi cluster-cluster kecil. Kemudian, kelompok-kelompok kecil ini secara bertahap menjauh dari tempat penampungan. Mungkin itu berarti emosi masyarakat sudah tidak tegang lagi saat ini.



Gambar 7.8. Karakteristik perubahan lintasan spatio-temporal masyarakat. Diagram urutan ini menggambarkan bagaimana massa berpindah dalam periode waktu kecil yang berbeda setelah gempa. Diantaranya, gambar (a) menunjukkan lintasan perubahan masyarakat dalam 10 menit setelah gempa. Tiga cluster terbentuk pada periode ini. Gambar (b) menunjukkan hubungan lokasi antara masing-masing cluster dan shelter pada sepuluh menit kedua. Gambar (c), (d) dan (e) menunjukkan bahwa semua cluster kecil membentuk cluster besar seiring berjalannya waktu dan memiliki populasi terbesar antara pukul 08:40 dan 09:00 seperti pada gambar (d). Gambar (f) dan (g) menunjukkan kerumunan secara bertahap menghilang dan meninggalkan shelter. Dari keseluruhan proses analisis, kami menyimpulkan bahwa: (1) Saat gempa terjadi, masyarakat bergegas ke tempat pengungsian dalam jangka waktu yang sangat singkat. Namun, apakah tempat penampungan ini ditata dengan baik? Kami melihat beberapa shelter tidak menampung banyak orang, atau bahkan tidak ada orang. Oleh karena itu, hasil analisis dapat digunakan sebagai acuan tata letak shelter yang rasional. (2) Karakteristik pengumpulan massa dan evakuasi dapat digunakan sebagai acuan efektif untuk membantu departemen pengurangan bencana dalam menghadapi keadaan darurat di masa depan.

Lebih lanjut, kami ingin mengetahui kategori emosi apa saja yang diungkapkan masyarakat selama periode ini dan bagaimana emosi tersebut berubah karena perpindahan penduduk secara besar-besaran dalam waktu singkat dapat menyebabkan beberapa kecelakaan yang tidak perlu seperti kemungkinan terinjak-injak karena panik. Oleh karena itu,

jika emosi masyarakat dalam proses ini dapat dipantau, hal ini akan membantu kita mengambil tindakan yang cepat dan efektif untuk meningkatkan efisiensi evakuasi dan mencegah kecelakaan. Tabel 7.1 menyajikan karakteristik emosi setiap periode waktu pada Gambar 7.8. Indikator pada tabel ini antara lain cluster yang dibentuk oleh kernel densitas clustering, kategori emosi dan kategori emosi utama yang terdapat pada setiap cluster, dan lain-lain.

Tabel 7.1. Distribusi karakteristik emosi dalam cluster yang berbeda dalam periode waktu yang berbeda.

Periode Waktu	Klaster	Kategori Emosi	Kategori Emosi Utama
08.02 Hingga 08.12 (Gambar 7.8a)	Cluster 1	Cemas	Cemas
	Cluster 2	Takut	Takut
	Cluster 3	Cemas, takut, marah	Takut
08.12 Hingga 08.22 (Gambar 7.8b)	Cluster 1	Cemas, takut, marah	Takut
	Cluster 2	Cemas, marah, takut	Takut
	Cluster 3	Marah, takut, positif	Takut
08.22 hingga 08.40 (Gambar c)	Cluster 1	Cemas, marah, Takut, sedih, netral, positif	Takut
08.40 Hingga 09.30 (Gambar 7.8d)	Cluster 1	Cemas, marah, Takut, netral, positif	Takut
09.30 Hingga 10.30 (Gambar 7.8e)	Cluster 1	Cemas, takut, Netral, cemas positif	Cemas, takut
10.30 Hingga 11.30 (Gambar 7.8f)	Cluster 1	Takut, positif, netral	Cemas
	Cluster 2	Netral, Sedih	Takut
11.30 Hingga 12.30 (Gambar 7.8g)	Cluster 1	Takut, netral positif, cemas, sedih	Netral, menyedihkan
	Cluster 2	Netral, Marah	Positif
	Cluster 3	Marah, Cemas, Netral	Marah
	Cluster 4	Netral, Positif	Cemas

Dari Tabel 7.1, kita dapat melihat bahwa emosi ketakutan mendominasi pada 150 menit pertama (8:02 hingga 10:30) setelah gempa bumi, diikuti oleh rasa cemas. Pada periode ini, masyarakat tidak siap menghadapi gempa bumi yang tidak terduga dan takut kehilangan nyawa akibat gempa tersebut. Di antara mereka, pada periode waktu pertama (08:02 hingga 08:12), masyarakat menyatakan kecemasan di cluster 1, seperti yang ditunjukkan pada Gambar 7.8a. Melalui konten teks terkait, kami menemukan bahwa mereka tidak mengetahui detail gempa saat ini (seperti lokasi pusat gempa, skala magnitudo, dll), sehingga mereka mengkhawatirkan keselamatan kerabat, teman, dan bahkan orang lain yang tidak mereka kenal. Emosi marah, netral, dan positif mulai muncul pada periode waktu kedua (08:12 hingga 08:22). Namun, jumlah emosi positif dan netral relatif kecil, dengan satu bagian positif dan dua bagian netral. Emosi marah pada periode ini terutama menunjukkan keengganan masyarakat terhadap gempa bumi. Emosi sedih mulai muncul pada periode waktu ketiga

(08:22 hingga 08:40) Orang-orang yang mengungkapkan emosi sedih terutama karena gempa ini mengingatkan mereka akan bencana mengerikan yang terjadi di Sichuan pada tahun 2008 (gempa yang terjadi di Wenchuan, Provinsi Sichuan, pada tanggal 12 Mei 2008, menyebabkan kerusakan besar). Dengan semakin banyaknya informasi rinci tentang gempa bumi yang tersedia, orang-orang mengekspresikan lebih banyak kategori emosi, dan alasan emosi ini juga berubah. Misalnya, pada periode keempat dan kelima, masyarakat mengungkapkan emosi takut dan cemas karena khawatir akan terjadi gempa susulan dalam waktu dekat. Hal ini juga menunjukkan bahwa dalam kurun waktu yang lama masyarakat masih berkumpul di dekat shelter, seperti terlihat pada Gambar 7.8c–e. Kita dapat melihat banyak orang mulai meninggalkan tempat penampungan pada Gambar 7.8f. Jika digabungkan dengan emosi yang diungkapkan orang-orang pada periode ini, kami menemukan bahwa emosi ketakutan tidak lagi dominan. Masyarakat mengungkapkan emosi sedihnya di klaster 2 pada periode keenam kalinya karena kesedihannya terhadap para korban di daerah yang paling parah terkena dampaknya. Pada periode waktu ketujuh, kategori emosi utama di setiap cluster berbeda-beda. Orang-orang mungkin sudah tenang saat ini. Bahkan emosi utama cluster 2 pada Gambar 7.8g adalah positif. Masyarakat menyampaikan doanya untuk daerah bencana.

Analisis emosi yang mendalam tidak hanya memberikan penjelasan lebih jauh mengenai penelusuran lintasan spatio-temporal manusia, namun juga memberikan rincian lebih lanjut mengenai bencana bagi departemen pengurangan bencana. Di satu sisi memberikan pemahaman mengenai kesadaran darurat masyarakat dan hukum pergerakan di wilayah studi. Di sisi lain, berdasarkan karakteristik lintasan emosional spatio-temporal, departemen pengurangan bencana dapat memberikan manajemen dan panduan yang tepat waktu untuk "simpul-simpul kunci" dalam proses ini untuk menghindari kecelakaan yang tidak terduga. Misalnya, kita dapat memberikan panduan yang efektif untuk area di mana emosi negatif sangat kuat, misalnya pada Gambar 8a,b, untuk menghindari kemungkinan terinjak-injak karena panik.

7.7 ANALISIS PERUBAHAN EMOSI PASCA BENCANA

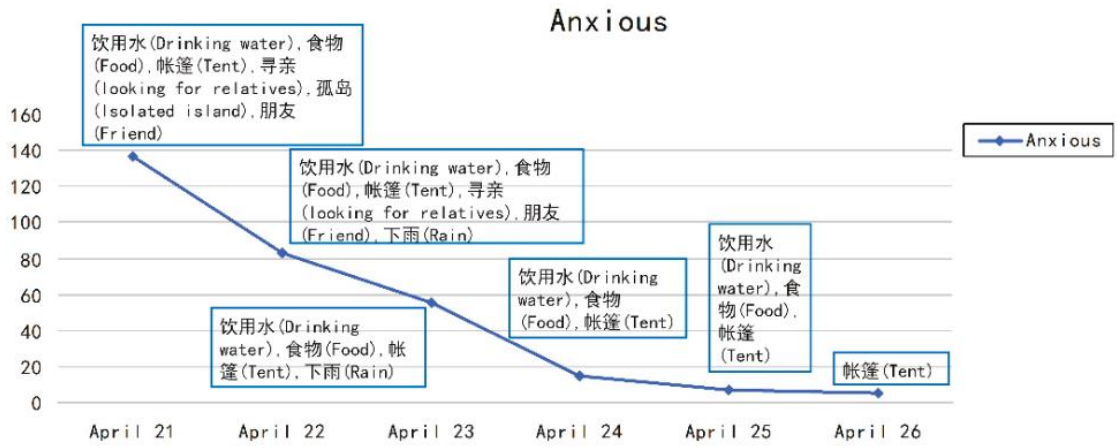
Bencana yang terjadi secara tiba-tiba ini mempunyai dampak jangka panjang bagi masyarakat. Dengan memantau dan menganalisis informasi emosional masyarakat yang terperinci, kita dapat menggali banyak informasi penting dari data besar terkait bencana. Informasi ini dapat membantu kita dengan cepat memahami kebutuhan dan masukan masyarakat, bahkan untuk beberapa masalah yang sulit ditemukan, seperti kesehatan mental. Hal ini sangat penting bagi kita untuk meningkatkan efisiensi tanggap darurat dan penyelamatan bencana. Pada bagian ini, kami mengambil Ya'an sebagai contoh dan menggunakan hari sebagai interval untuk memantau emosi publik dalam jangka waktu yang lebih lama dari perspektif makro. Sementara itu, kami menganalisis penyebab perubahan emosi masyarakat menggunakan ekstraksi kata panas. Hal ini dapat membantu kita dengan cepat memahami kekhawatiran masyarakat dari informasi emosional massa. Alat yang kami gunakan untuk mengekstrak kata-kata hangat berasal dari web (http://www.picdata.cn/).

Mengenai emosi positif, seperti yang ditunjukkan pada Gambar 7.9, pada hari kedua (21 April) setelah bencana, kami memasukkan kata-kata panas "感动 (terharu)" dan 感激 (bersyukur)" ke dalam teks yang sesuai. Kami menemukan bahwa masyarakat mengungkapkan rasa terima kasih mereka terutama kepada petugas penyelamat profesional, seperti tentara. Jumlah sukarelawan saat ini lebih sedikit. Namun seiring berjalannya waktu, semakin banyak relawan yang secara spontan bergabung dalam upaya penyelamatan, terutama pada hari ketiga dan keempat. Karena saat ini "志愿者(sukarelawan)" lebih sering muncul. Tim penyelamat sipil secara bertahap tiba di lokasi bencana sekitar hari keempat (23 April) setelah bencana. Kata-kata hangat "爱心 (cinta)" dan "物资 (persediaan bantuan)" menunjukkan bahwa masyarakat di daerah yang terkena bencana berterima kasih atas bantuan bantuan spontan non-pemerintah. Berdasarkan perubahan emosi positif masyarakat, kita dapat memahami proses umum upaya penyelamatan.

Gambar 7.9. Diagram urutan emosi positif (kata-kata di kotak teks adalah kata-kata hangat yang terkait dengan emosi ini dalam periode waktu yang sesuai).

Terkait dengan kecemasan, seperti terlihat pada Gambar 7.10, seiring berjalannya waktu, kecemasan masyarakat berangsur-angsur berkurang. Pada hari kedua setelah gempa, kecemasan semakin bertambah. Penyebabnya karena: (1) Jalan terputus dan beberapa daerah terisolasi. Kami menggunakan kata-kata populer "中断 (interupsi)" dan "救援 (penyelamatan)" dalam teks mikro-blog asli untuk mendapatkan informasi rinci. Kami menemukan bahwa "上里镇 (kota Shangli)," "中里镇 (kota Zhongli)," "下里镇 (kota Xiali)," dan "碧峰峡 (kota Bifengxia)" terisolasi dari dunia luar dan membutuhkan penyelamatan mendesak setelah gempa bumi. (2) Beberapa daerah sangat membutuhkan pasokan bantuan, dan daerah tersebut terkena dampak cuaca buruk. Misalnya, melalui kata-kata hangat, kami menemukan ada beberapa teks mikro-blog yang mengatakan bahwa "Desa Wangjia di kota

Longmen kekurangan air, makanan, obat-obatan, dan tenda". (3) Beberapa masyarakat menyatakan cemas karena tidak dapat menghubungi kerabat dan temannya pasca gempa. Pada tanggal 21 April hingga 26 April, kami menemukan kekhawatiran masyarakat terutama disebabkan oleh kurangnya pasokan dan cuaca buruk. Selain itu, seiring berjalannya waktu, kebutuhan masyarakat akan bantuan terutama berupa tenda, terutama pada tanggal 26 April. Kombinasi konten mikro-blog dan informasi lokasi memungkinkan dilakukannya rencana penyelamatan yang lebih tepat.

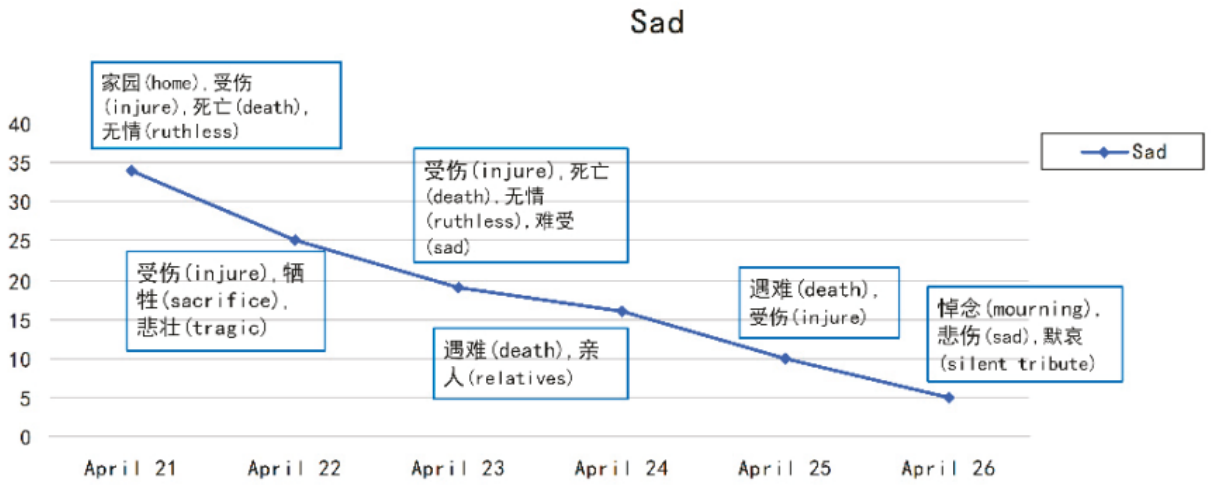


Gambar 7.10. Diagram urutan kecemasan.

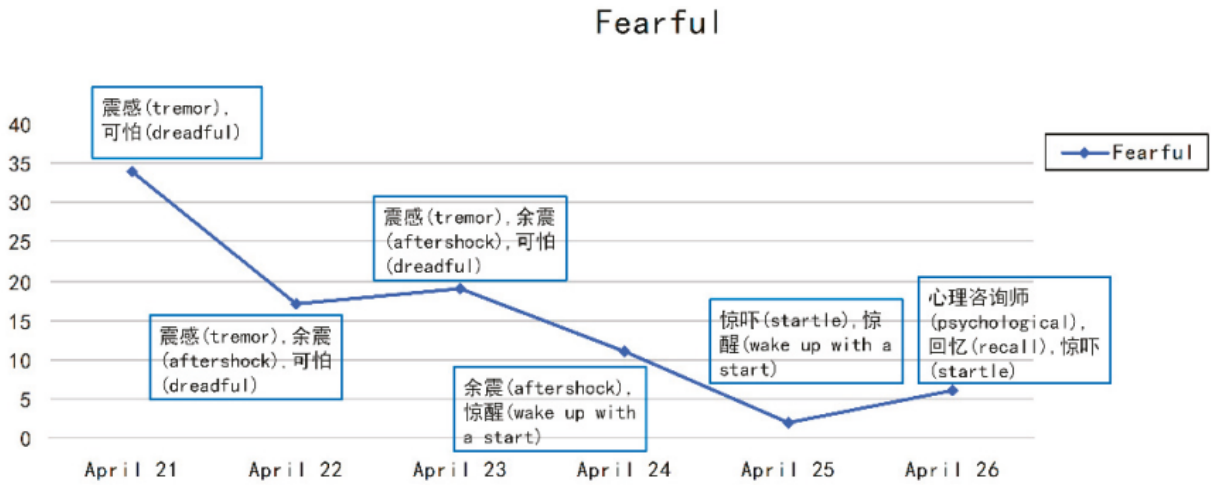
Seperti yang ditunjukkan pada Gambar 7.11, dari tanggal 21 April hingga 23 April, masyarakat sebagian besar mengungkapkan kemarahannya karena keengganan mereka terhadap gempa bumi dan karena beberapa penipuan internet. Dengan menggabungkan konten mikro-blog yang sesuai, kami menemukan bahwa beberapa penipuan internet diungkap oleh pengguna mikro-blog, seperti "Mengerikan. Beberapa penjahat melakukan penipuan di balik kedok gempa. Mohon diperhatikan nomor telepon ini: xxx.". Pesan-pesan kemarahan ini digunakan untuk membantu orang-orang agar terhindar dari rumor (terutama bagi orang-orang yang cemas). Setelah tanggal 23 April, kemarahan masyarakat terutama disebabkan oleh keengganan terhadap bencana tersebut.

Gambar 7.11. Diagram urutan kemarahan.

Seperti terlihat pada Gambar 12, pada tanggal 21 April hingga 25 April, kesedihan terjadi akibat hancurnya rumah dan meninggalnya kerabat atau teman. Contoh teks mikro-blog yang sesuai adalah: “Di mana rumahnya? Dimana ruang kelasnya? Kemarin? Saya bukan seorang yatim piatu kemarin” dan “Ini adalah pemandangan yang benar-benar hancur dan kapan kita bisa membangun kembali tanah air kita?” Pada tanggal 26 April, banyak kegiatan berkabung yang dilakukan oleh lembaga resmi dan non-pemerintah, sehingga menjadi alasan masyarakat mengungkapkan kesedihannya. Selama proses penyelamatan gempa bumi, organisasi bantuan bencana dapat mengirimkan bantuan psikologis ke tempat-tempat di mana kesedihan sangat mendalam, sebagaimana ditentukan oleh lokasi informasi mikro-blog terkait.



Gambar 7.12. Diagram urutan kesedihan.



Gambar 7.13. Diagram urutan ketakutan.

Dari segi ketakutan, seperti terlihat pada Gambar 7.13, pada tanggal 22 April hingga 24 April terjadi beberapa kali gempa susulan di Ya’an yang berdampak besar terhadap kehidupan masyarakat. Banyak kata-kata panas yang diamati, seperti “余震 (gempa susulan)” dan “可怕 (mengerikan)”. Namun, antara tanggal 24 April dan 26 April, terutama pada tanggal

26 April, tiba-tiba ada peningkatan rasa takut, dan kata-kata panas tersebut terutama mencakup "心理咨询师(konselor psikologis)," "回忆 (ingat)", dan "惊吓(mengejutkan). Kami menggunakan kata-kata panas ini untuk mendapatkan mikro-blog asli dan melihat beberapa orang mengatakan bahwa: "Seorang guru di Kota Zhongli melaporkan bahwa seorang gadis takut dengan suara keras dan terus makan. Dia bilang dia akan takut jika dia tidak makan. Jadi guru ini berharap departemen pengurangan bencana dapat mengirimkan konselor psikologis untuk membantu gadis itu" dan "Adikku bilang selama ada guntur dan kilat di Ya'an, dia sangat ketakutan! Meminta bantuan!" Oleh karena itu, departemen bantuan harus memberikan bantuan kepada orang-orang ini.

Kita bisa terus mengeksplorasi area lain dengan cara yang sama. Hal ini dapat membantu kita memahami secara akurat reaksi masyarakat terhadap kemajuan mitigasi bencana. Lebih lanjut, hasil analisis dapat membantu departemen penyelamatan untuk mengoptimalkan strategi penyelamatan dan meningkatkan efisiensi penyelamatan.

7.8 EVALUASI AKURASI KLASIFIKASI EMOSI

Korpus Eksperimental

Dalam eksperimen klasifikasi emosi, pertama-tama kami membuat anotasi korpus secara manual berdasarkan enam kategori emosi. Dalam korpus ini, setiap kategori emosi berisi 1000 sampel teks. Diantaranya, 800 sampel teks dipilih sebagai korpus pelatihan dan 200 dipilih sebagai korpus pengujian dari masing-masing kategori emosi.

Lingkungan Eksperimental

Untuk meningkatkan keakuratan klasifikasi emosi, kami menerjemahkan karakter dan simbol khusus dalam teks ke dalam kata-kata berbahasa Mandarin. Kami mengintegrasikan kerangka kerja word2vec dan NLPIR-ICTCLAS (http://ictclas.nlpir.org) ke dalam kerangka algoritmik kami untuk membantu pemrosesan teks ini. Terakhir, kami membangun jaringan saraf konvolusional berdasarkan aliran tensor, dan mengoptimalkan parameter model untuk mencapai hasil terbaik. Dalam proses ini, kami menetapkan dimensi vektor kata sebagai 200, jumlah kernel konvolusi adalah 3 dan ukurannya adalah 3, 4, dan 5. Ukuran max pooling adalah 4, proporsi regularisasi dropout adalah 0,3, dan langkahnya adalah 1.

Hasil Eksperimen dan Perbandingan Akurasi

Kami memverifikasi keakuratan algoritme berdasarkan presisi (P), perolehan (R), dan indeks evaluasi komprehensif (F-1). Rumusnya ditunjukkan di bawah ini:

Persamaan 5
$$P = \frac{N\_Correct}{N\_Correct + N\_False}$$

Persamaan 6
$$R = \frac{N\_Correct}{N\_Category}$$

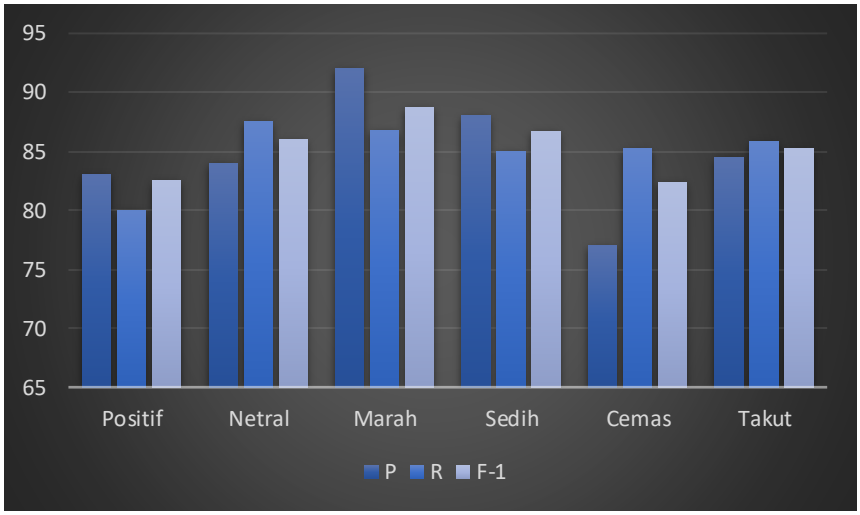
Persamaan 7
$$F - 1 = \frac{2 \times P \times R}{P + R}$$

N_Correct mewakili jumlah teks yang diklasifikasikan dengan benar ke dalam satu kategori, N_False mewakili jumlah teks yang salah diklasifikasikan ke dalam kategori ini, dan N_Category mewakili jumlah teks yang termasuk dalam kategori ini dalam korpus pengujian.

Tabel 7.2 dan Gambar 7.14 menunjukkan keakuratan model CNN dalam klasifikasi emosi yang terperinci. Skor indeks evaluasi komprehensif untuk setiap kategori semuanya di atas 81%, yang memenuhi persyaratan eksperimental. Selain itu, dalam bab ini, kami juga mempertimbangkan penggunaan bahasa gaul, kata kunci internet, serta karakter dan simbol khusus untuk meningkatkan performa model.

Tabel 7.2. Evaluasi akurasi klasifikasi emosi positif.

Kategori Emosional	Presisi (P)	Recall (R)	Evaluasi Komprehensif Inderks (F-1)
Positif	82,25%	80,00%	82,54%
Netral	84,21%	87,91%	86,02%
Marah	91,57%	86,36%	88,89%
Sedih	88,54%	85,00%	86,73%
Cemas	78,47%	85,27%	81,77%
Takut	84,69%	85,57%	85,13%



Gambar 7.14. Akurasi klasifikasi berbagai emosi.

Deskripsi Data dengan Informasi Alamat

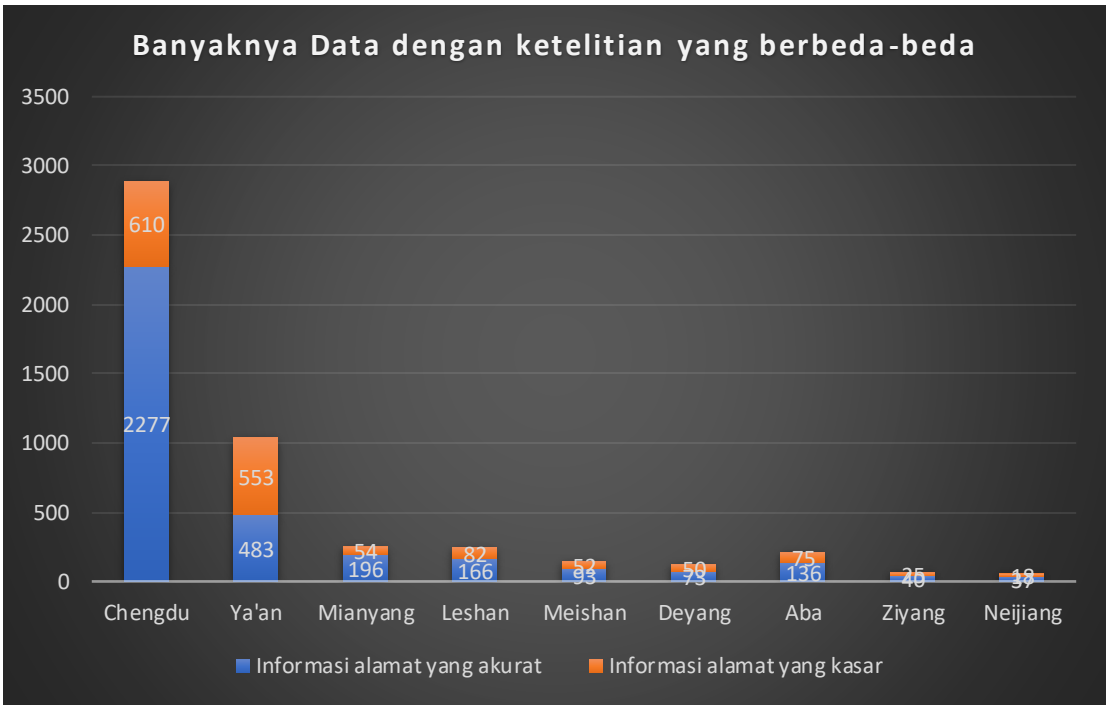
Jumlah teks dalam kumpulan data bab ini adalah 39341. Namun tidak semua teks memuat informasi alamat. Semua informasi alamat dalam tulisan ini dapat dibagi menjadi dua kategori, antara lain informasi alamat kasar dan informasi alamat akurat. Diantaranya, informasi alamat kasar hanya dapat mewakili desa dan kota kecil, bahkan distrik dan kabupaten, seperti “Kabupaten Lushan, Kota Ya'an” dan “Distrik Wuhou, Kota Chengdu,” dll. Informasi alamat yang akurat dapat mewakili jalan dan entitas geografis, seperti “Jalan Sishengci Utara,” “Kampus Wangjiang Universitas Sichuan,” dll. Tabel 7.3 dan Gambar 7.15

menggambarkan proporsi dan jumlah data dengan keakuratan berbeda di berbagai kota. Diantaranya rumus menghitung proporsinya adalah sebagai berikut: (Persamaan 8)

$$\text{Proporsi} = \frac{\text{Jumlah teks tertentu di kota}}{\text{Jumlah semua teks di kota}}$$

Tabel 7.3. Proporsi data dengan akurasi berbeda di berbagai kota.

Kota	Proporsi Data Dengan Informasi Alamat Yang Akurat	Proporsi Data Dengan Informasi Alamat Kasar	Total
CHENGDU	8,08%	2,16%	10,24%
YA'AN	12,18%	13,91%	26,09%
MIANYANG	9,45%	2,60%	12,05%
LESHAN	9,81%	4,85%	14,66%
MEISHA	13,30%	7,44%	20,74%
DEYANG	8,71%	5,79%	14.50%
ABA	11,29%	23,51%	34,80%
ZIYANG	5,10%	3,10%	8.20%
NEIJIAN	4,79%	2,33%	7.12%



Gambar 7.15. Perbandingan jumlah informasi alamat dengan akurasi yang berbeda-beda di setiap kota.

Evaluasi Proses dan Hasil Eksperimental

Pada pembahasan sebelumnya, kami menggunakan data distribusi kepadatan penduduk yang disediakan oleh GHSL (Global Human Settlement Layer) untuk membantu menilai populasi yang terkena dampak. Skala peta yang kami gunakan relatif kecil. Oleh karena itu, kami menilai semua data baik informasi alamat akurat maupun informasi alamat kasar dapat digunakan. Diantaranya, data di Aba dan Ya'an lebih mencerminkan situasi sebenarnya

yang diungkapkan oleh media sosial di wilayah tersebut karena meskipun volume data media sosial di kedua kota ini kecil, namun data dengan informasi alamat menyumbang proporsi yang lebih besar; masing-masing mencapai 34,8% dan 26,09%. Mengingat kepadatan penduduk dan lokasi episentrum (pusat gempa terjadi di Ya'an),

7.10 RINGKASAN

Ketika bencana terjadi, media sosial dapat memberikan sejumlah besar informasi geografis penting terkait bencana kepada departemen pengurangan bencana hampir secara real-time. Dalam bab ini, kami menganggap informasi emosional publik yang diperoleh dari media sosial sebagai atribut informasi geografis untuk membantu mitigasi bencana. Dalam proses mengekstraksi informasi emosional, kami menganalisis sepenuhnya karakteristik media sosial Tiongkok dan memilih algoritma yang sesuai (model jaringan saraf konvolusi). Sementara itu, banyaknya karakter dan simbol khusus dengan ciri-ciri emosional yang terdapat di media sosial juga dinilai dapat meningkatkan akurasi klasifikasi.

Untuk memverifikasi efektivitas metode dalam bab ini dalam mitigasi bencana, kami menggunakan gempa bumi berkekuatan 7,0 yang terjadi pada tanggal 20 April 2013, di Kota Ya'an, Provinsi Sichuan, Tiongkok, sebagai studi kasus. Kami mengklasifikasikan teks media sosial terkait wilayah yang terkena dampak gempa ke dalam enam kategori emosi berbeda. Kemudian, dengan bantuan perangkat lunak GIS dan data informasi geografis tradisional lainnya (data sebaran kepadatan penduduk dan data tempat tinggal), kami mengeksplorasi peran informasi emosional masyarakat yang berguna dalam pengurangan bencana. Hasilnya menunjukkan bahwa emosi masyarakat yang mendalam dapat memberikan dukungan data yang lebih kuat bagi departemen pengurangan bencana untuk mengoptimalkan strategi penyelamatan dan meningkatkan efisiensi penyelamatan.

Meskipun media sosial berperan penting dalam membantu mitigasi bencana, media sosial juga mempunyai beberapa keterbatasan. (1) Data media sosial tidak terdistribusi secara merata. Daerah yang ekonominya maju dan padat penduduknya cenderung memiliki lebih banyak pengguna mikroblog Sina. Chengdu memiliki data mikro-blog Sina terbanyak dan data ini lebih terkonsentrasi di wilayah perkotaan, namun kota ini bukanlah kota yang paling parah terkena dampaknya. Oleh karena itu, pada penelitian selanjutnya, diperlukan juga lebih banyak data yang mencakup sumber lain untuk melengkapi data media sosial, seperti data gambar, data GPS kendaraan, dll. (2) Tidak semua pengguna media sosial bersedia membagikan informasi lokasinya. Pada dataset yang digunakan dalam bab ini, proporsi teks dengan informasi lokasi sangat kecil. Hal ini membatasi penggunaan beberapa metode analisis spatio-temporal. Namun, kami menemukan bahwa ada banyak entitas yang diberi nama geografis dalam teks dan banyak di antaranya yang dapat mengetahui lokasi pengguna. Oleh karena itu, diperlukan metode yang efektif untuk secara otomatis mengekstraksi entitas yang diberi nama geografis ini untuk melengkapi kekurangan informasi lokasi geografis di media sosial.

BAB 8
METODE PEMBUATAN JALAN BERDASARKAN DATA NAVIGASI SELULER

Dengan pesatnya perkembangan kota, informasi geografis blok perkotaan juga berubah dengan cepat. Namun, metode tradisional dalam memperbarui data jalan tidak dapat mengikuti perkembangan ini karena memerlukan keahlian profesional tingkat tinggi untuk pengoperasiannya dan sangat memakan waktu. Dalam bab ini, kami mengembangkan metode baru untuk mengekstraksi jalan raya yang hilang dengan merekonstruksi topologi jalan dari data lintasan navigasi seluler yang besar. Tiga langkah utama tersebut meliputi penyaringan data lintasan navigasi asli, mengekstraksi garis tengah jalan dari titik navigasi, dan menetapkan topologi jalan yang ada. Pertama, data pejalan kaki dan pengemudi di jalan yang ada dihapus dari data mentah. Kedua, garis tengah jalan blok kota diekstraksi menggunakan metode RSC (ring-stepping clustering) yang diusulkan di sini. Terakhir, topologi jalan yang hilang dan hubungan antara jalan yang hilang dan jalan yang ada dibangun. Sebuah blok perkotaan yang kompleks dengan luas 5,76 kilometer persegi dipilih sebagai wilayah studi kasus. Validitas metode yang diusulkan diverifikasi menggunakan kumpulan data yang terdiri dari data lintasan navigasi seluler selama lima hari. Hasil percobaan menunjukkan rata-rata kesalahan absolut panjang garis tengah yang dihasilkan adalah 1,84 m. Analisis komparatif dengan metode ekstraksi jalan lain yang ada menunjukkan bahwa kinerja F-score dari metode yang diusulkan jauh lebih baik dibandingkan metode sebelumnya.

8.1 PENDAHULUAN

Dengan pesatnya pembangunan konstruksi jalan di wilayah perkotaan dan pedesaan, perubahan jalan yang diakibatkannya mengakibatkan lambatnya pembaruan data jalan yang tidak sesuai dengan situasi saat ini serta memiliki integritas dan akurasi yang rendah. Teknologi tradisional yang digunakan untuk mendeteksi dan memperbarui jalan yang hilang, seperti survei profesional, perampingan peta, interpretasi gambar penginderaan jauh, dll., lebih mahal, memerlukan siklus pembaruan yang lebih lama, lebih rumit dalam pemrosesan data, dan tidak dapat dengan mudah beradaptasi dengan kebutuhan jalan. perkembangan kota yang pesat.

Mendeteksi jalan baru dan jalan yang hilang pada jaringan jalan yang ada telah menjadi perhatian umum di bidang manajemen perkotaan, transportasi cerdas, dan teknologi tanpa pengemudi. Citra, lintasan GNSS (sistem satelit navigasi global), dan fusi data multisumber adalah beberapa sumber data utama yang digunakan untuk pembaruan dan perbaikan data jaringan jalan GIS (sistem informasi geografis) yang hilang. Dengan perkembangan teknologi seperti komunikasi nirkabel, Big Data, dan komputasi awan, lintasan navigasi secara bertahap menjadi sumber data utama untuk pembaruan jalan perkotaan. Lintasan pemosisian dalam kendaraan yang masif memiliki karakteristik data yang besar, banyak sumber data, dan

struktur yang heterogen. Penggunaan lintasan berbasis VGI (informasi geografi sukarela), seperti yang dikumpulkan oleh ponsel pintar, dan perangkat pintar yang dipasang di kendaraan atau dipegang oleh pejalan kaki, untuk memperbarui data jalan baru-baru ini telah dicapai. hasil yang substansial.

Dalam buku ini, diusulkan sebuah metode baru untuk mengekstraksi jalan yang hilang di blok kota dari data lintasan navigasi bergerak yang besar, yang utamanya terdiri dari langkah-langkah berikut: (1) Informasi yang tidak berguna dari pengguna pejalan kaki mengenai jalan-jalan perkotaan yang ada disaring dari data; (2) Setelah pra-pemrosesan, garis tengah jalan diekstraksi menggunakan metode RSC (ring-stepping clustering) yang diusulkan; (3) Topologi jalan setiap blok kemudian ditetapkan, dan hubungan antara jalan perkotaan yang ada dan jalan yang hilang di blok tersebut dibangun. Dibandingkan dengan metode tradisional, metode yang diusulkan memiliki keuntungan sebagai berikut: Pertama, dibandingkan dengan kendaraan khusus yang dilengkapi dengan perangkat pemetaan, seluruh proses dapat mencapai otomatisasi tingkat tinggi dan jarang memerlukan operasi manual, dan investasi perangkat survei juga jauh lebih kecil; kedua, dibandingkan dengan ekstraksi citra penginderaan jauh, pengaruh ekstraksi pohon yang terdapat di mana-mana di jalan raya pada blok-blok kota terhadap ekstraksi jalan raya akan jauh lebih kecil, dan periode kunjungan ulang satelit tidak akan lagi menjadi masalah; terakhir, dibandingkan dengan metode ekstraksi jalan yang ada menggunakan data lintasan GNSS, meskipun kompleksitas komputasi dari metode yang diusulkan (Θ(n2)) lebih tinggi dibandingkan metode pada (Θ(n)), skor F − garis tengah jalan yang dihasilkan jauh lebih tinggi.

Sebuah studi kasus dengan 9.944.710 titik lintasan GNSS di area seluas 5,76 kilometer persegi dipilih untuk memverifikasi kelayakan metode yang diusulkan. Hasilnya menunjukkan bahwa jaringan jalan yang diekstraksi sudah sesuai dengan jaringan sebenarnya.

8.2 PEMUTAKHIRAN DATA GEOMETRI JALAN

Pemilihan titik lintasan GNSS yang tidak terletak pada jalan yang sudah ada biasanya merupakan tahap penting dalam pemetaan jalan untuk memperbarui jaringan jalan dan menyempurnakan geometri ruas atau persimpangan jalan. Kemudian, algoritma, strategi, dan metode yang berbeda untuk mendeteksi jalan baru, perluasan, dan hilangnya jalan yang ada diusulkan berdasarkan titik lintasan outlier.

Mengingat lintasan—atau titik-titik—hanya muncul di jalan raya, pengelompokan titik-titik GNSS adalah metode yang paling umum digunakan untuk memperbarui data geometrik suatu jalan. Misalnya, K-Means telah digunakan untuk mengelompokkan sejumlah besar titik GNSS pada posisi tertentu di jalan untuk mengidentifikasi pusat jalan. Metode semacam ini dilanjutkan dengan memasukkan titik atau lintasan GNSS dan menentukan titik awal untuk mengelompokkan pusat jalan. Kemudian, setelah dilakukan iterasi pada interval jarak tertentu, diperoleh informasi geometri jaringan jalan.

Metode penggabungan jejak biasanya digunakan untuk ekstraksi jalan dari lintasan GNSS yang masif. Dengan menggunakan iterasi setiap lintasan GNSS, tepi lintasan mentah ditambahkan ke peta jalan saat ini sesuai dengan hasil pencocokan peta. Tepian peta jalan

saat ini diberi bobot untuk menggambarkan kejadian yang berulang dan kemungkinan keberadaan suatu jalan. Tepi yang memiliki bobot lebih rendah dihilangkan, dan tepi lainnya dianggap sebagai jalan yang baru ditemukan.

Sejumlah besar titik GNSS yang tidak berurutan menunjukkan ciri-ciri geometris yang tersebar di sepanjang jalan. Berdasarkan fenomena ini, peneliti mengusulkan metode estimasi kepadatan kernel (KDE) dan mengekstraksi area titik GNSS yang paling padat, ditambah dengan ambang batas tertentu, untuk mencapai tujuan ekstraksi batas jalan dan ekstraksi kerangka jalan. Keuntungan metode ini adalah seiring bertambahnya jumlah sampel, keluarannya akan lebih andal dan kuat. Namun, jika jumlah sampel tidak mencukupi, hasilnya sering kali mempunyai penyimpangan yang besar.

Baru-baru ini, total kuadrat terkecil dan metode deteksi titik balik diusulkan untuk mensegmentasi dan mengelompokkan data lintasan GNSS. Setelah dilakukan segmentasi dan pengelompokan, ditentukan posisi simpang dan ruas jalan dengan menggunakan fenomena perpotongan dan persilangan kelompok yang berbeda. Algoritme pembengkokan waktu dinamis (DTW) kemudian digunakan untuk menyelaraskan segmen jalan dari matriks koneksi antar persimpangan.

Namun, ekstraksi jaringan jalan raya masih menghadapi kesulitan di beberapa tempat. Misalnya, ketika jalan ditutupi oleh jembatan atau ada jalan bawah tanah di bawah jalan, titik-titik GNSS akan sangat kacau karena posisi GNSS ponsel saat ini tidak bagus di ketinggian; Titik-titik GNSS bisa sangat kacau, sehingga menyulitkan untuk mendapatkan bagian jalan ini dengan menggunakan metode yang ada.

Pemutakhiran Data Atribut Jalan

Kegunaan lain dari lintasan pergerakan adalah memperbarui informasi atribut jaringan jalan. Perubahan informasi atribut jalan terutama terbagi dalam delapan kategori: arah, batas kecepatan, jumlah jalur, akses, kecepatan rata-rata, kemacetan, kepentingan, dan offset geometrik. Winden dkk. mengusulkan algoritma pohon keputusan untuk menurunkan delapan atribut di atas untuk peta jalan terbuka (OSM). Hasilnya menunjukkan bahwa suatu jalan merupakan jalan satu atau dua arah, diklasifikasikan dengan akurasi 99% dan akurasi batas kecepatan jalan 69%.

Model kecepatan lintasan dari garis tengah jalan sebagai distribusi Gaussian, dan kemudian mengekstraksi atribut arah dan pembatasan belokan untuk peta jalan seperti OSM. Referensi menggunakan metode probabilistik untuk memperoleh jumlah jalur lalu lintas dari jalur GNSS dengan memasang model campuran Gaussian (GMM) pada persimpangan antara jalur GNSS dan garis pengambilan sampel yang tegak lurus terhadap garis tengah jalan. Teknik data-mining untuk mengekstrak nama dan kelas jalan dengan mengintegrasikan lintasan pergerakan dan data geotag dari media sosial dengan metode support vector machine (SVM). Jaringan saraf dalam (DNN) yang terhubung sepenuhnya untuk secara otomatis mengekstrak fitur dalam dan mengklasifikasikan lintasan berdasarkan mode transportasi. Kerangka kerja menggunakan rekayasa fitur dan penghilangan kebisingan untuk mengklasifikasikan lintasan pergerakan ke dalam moda transportasi umum seperti taksi, mobil, kereta api, kereta bawah tanah, jalan kaki, pesawat terbang, perahu, sepeda, lari, sepeda motor, dan bus.

Moda transportasi yang terdeteksi pada setiap lintasan pergerakan dapat digunakan untuk memperbarui atribut peta jalan. Algoritma pencocokan peta yang kuat untuk memastikan bahwa setiap titik ditetapkan ke peta jalan saat ini. Kemudian, persimpangan yang hilang, pembatasan belokan, dan penutupan jalan dideteksi dan diperbarui. Mengingat OSM telah menjadi cara umum bagi relawan untuk menggambar peta, data jalan baru dengan menganalisis data OSM. Mereka kemudian mengadopsi metode buffering progresif untuk memperbarui jalan terbaru dalam data OSM dengan jalan dari sumber data lain, termasuk geometri dan atributnya.

8.3. PENGEMBANGAN METODOLOGI

Metode yang diusulkan dapat dibagi menjadi tiga bagian utama (Gambar 8.1). Bagian pertama adalah pemfilteran data. Pada langkah ini, dua jenis data, termasuk titik navigasi yang terletak di jalan perkotaan yang ada dan catatan yang dihasilkan oleh pejalan kaki, disaring. Data navigasi yang tersisa dianggap sebagai titik navigasi yang berkaitan dengan mengemudi di jalan yang hilang di blok kota.

Kemudian, algoritma berbasis clustering bernama RSC diusulkan untuk mengekstraksi garis tengah jalan menggunakan titik navigasi yang dicadangkan. Ini adalah bagian kedua dari metode yang diusulkan. Setelah itu, garis tengah jalan yang hilang ditentukan. Akhirnya, topologi jalan yang hilang dan hubungannya dengan jalan yang ada dapat ditentukan.

Penyaringan Data

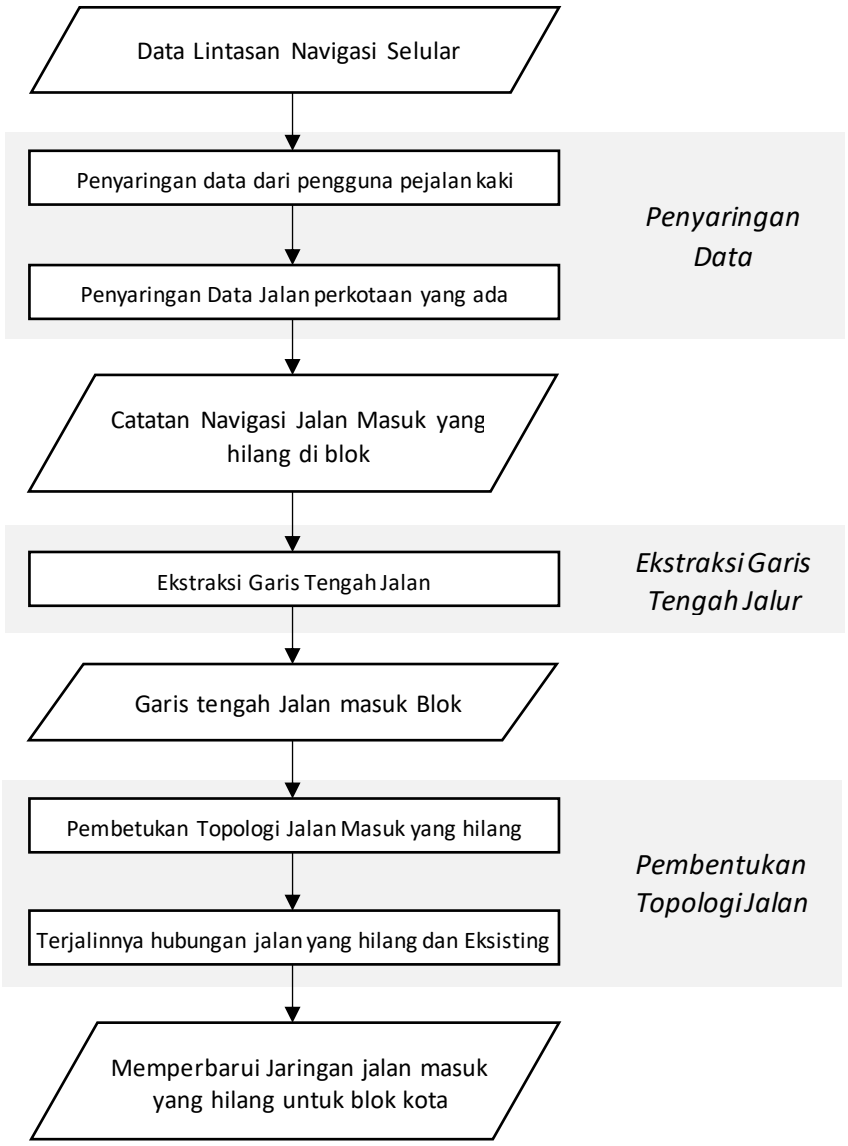
Data navigasi seluler dikumpulkan selama perangkat lunak navigasi aktif. Untuk mengurangi durasi penghitungan, titik data yang terletak di jalan eksisting dan dihasilkan oleh pejalan kaki harus disaring.

Pada subbagian ini, dua langkah utama dilakukan untuk menyaring data asli. Langkah pertama adalah menyaring data dari pengguna pejalan kaki berdasarkan kecepatan yang ditunjukkan oleh catatan. Kemudian, penyaringan dilakukan untuk memisahkan titik-titik GNSS berdasarkan jalan perkotaan yang ada melalui analisis overlay.

Pertama, informasi kecepatan titik GNSS merupakan indikator yang paling tepat untuk membedakan antara pejalan kaki dan pengguna kendaraan. Berdasarkan penelitian sebelumnya, kecepatan rata-rata berjalan kaki dan mengemudi masing-masing kira-kira 4,2 dan 30 km/jam. Oleh karena itu, dalam penelitian ini, 5,0 km/jam dijadikan ambang batas untuk memisahkan pengguna pejalan kaki dengan pengguna navigasi lainnya. Dengan kata lain, titik-titik GNSS dengan kecepatan di bawah 5,0 km/jam tersebut dianggap sebagai pengguna pejalan kaki dan kemudian dihapus.

Data navigasi yang dihasilkan oleh pergerakan pengguna sebagian berada pada jalan eksisting dan sisanya pada jalan yang hilang. Menurut penelitian kami saat ini, titik data navigasi di jalan yang ada mencakup lebih dari 90% dari seluruh titik dalam kumpulan data yang digunakan. Artinya, titik data navigasi pada jalan di dalam blok kota hanya menyumbang 10% dari total volume data. Oleh karena itu, menyaring titik data dari data asli jalan yang ada akan meningkatkan efisiensi metode ini karena data tersebut mencakup sebagian besar data.

Studi sebelumnya telah dilakukan untuk menyaring titik-titik posisi pada jalan perkotaan yang ada, seperti analisis overlay antara lapisan titik navigasi dan lapisan jalan perkotaan yang ada. Analisis vektor-raster juga dapat digunakan untuk pemfilteran. Dalam buku ini, kami mengadopsi metode untuk mengubah jalan yang ada dan titik navigasi ke dalam format raster, sehingga komputasi pemfilteran dapat dipercepat.



Gambar 8.1. Diagram alir metode yang diusulkan.

Ekstraksi Garis Tengah Jalan

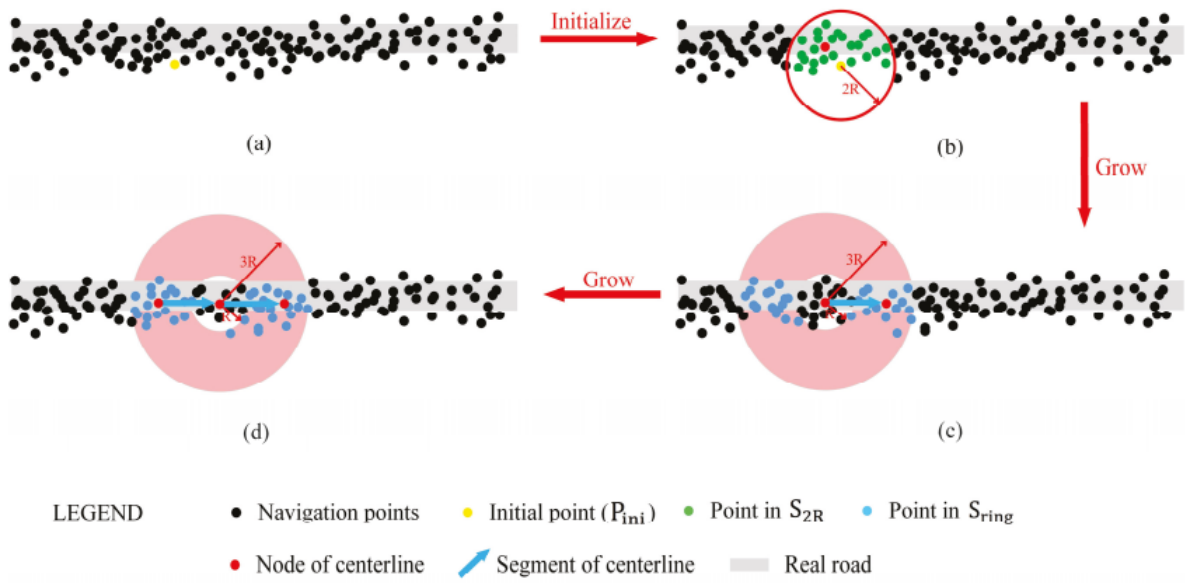
Setelah memfilter data mentah, titik navigasi yang lebih murni dari jalan-jalan yang hilang di blok kota akhirnya dicadangkan. Untuk mengekstraksi jaringan jalan dari titik-titik tersebut, algoritma berbasis clustering yang disebut RSC diusulkan pada subbagian ini.

Ekstraksi Garis Tengah melalui RSC

Pada subbagian ini, RSC digunakan untuk mengekstraksi garis tengah jalan yang hilang. Gambar 8.2 menunjukkan empat langkah utama RSC. Titik-titik hitam menunjukkan titik navigasi, yang dicadangkan setelah pemfilteran data,.

Mengingat titik cadangan GNSS terletak di jalan yang hilang, maka titik awal (Pini) dipilih secara acak dari titik cadangan (lihat titik kuning pada Gambar 8.2a). Kemudian digunakan parameter radius (R) untuk memilih titik-titik yang dicakup oleh cincin berjari-jari 2R (S2R) yang meliputi lingkaran merah dan titik hijau pada Gambar 8.2b. R adalah parameter radius lingkar, yang kira-kira merupakan lebar rata-rata jalan di wilayah ini. Setelah itu, posisi kepadatan tertinggi dapat dihitung dengan menggunakan titik-titik berwarna hijau dan selanjutnya dapat dianggap sebagai simpul dari garis tengah (lihat titik merah pada Gambar 8.2b).

Setelah titik tengah jalan awal ditemukan, sebuah cincin dengan diameter dalam R dan jari-jari luar 3R ditentukan untuk mencari simpul garis tengah jalan berikutnya. Titik-titik yang jatuh pada ring diangkat membentuk himpunan titik Sring (lihat titik biru pada Gambar 8.2c).

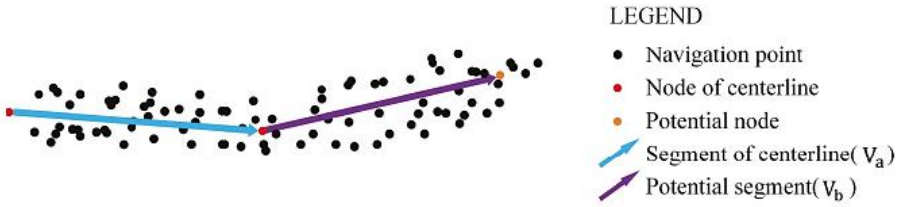


Gambar 8.2. Langkah-langkah utama RSC (ring-stepping clustering) adalah sebagai berikut: (a) titik GNSS (sistem satelit navigasi global) acak dipilih sebagai titik awal, (b) node awal dipilih dari kumpulan titik S2R dari titik awal, dan (c,d) node berikutnya dipilih dari himpunan titik Sring dari node saat ini.

Kepadatan setiap titik di Sring juga dihitung untuk mencari titik kepadatan tertinggi. Namun, sebelum memutuskan simpul berikutnya, setiap titik di Sring dihitung untuk menentukan apakah titik tersebut memenuhi putaran U. Seperti terlihat pada Gambar 8.3 dan Persamaan (1), nilai V dihitung dari segmen potensial (Vb) dan garis tengah segmen eksisting (Va). (Persamaan 1)

$$V = V_a \cdot V_b$$

Jika V < 0, maka node potensial tidak akan dipilih sebagai node berikutnya.



Gambar 8.3. Menentukan putaran U.

Pseudocode dari algoritma RSC yang diusulkan disajikan dalam Algoritma 1. Selama pemrosesan data, pohon k–d diperkenalkan sebagai indeks titik yang ditetapkan untuk percepatan. Kompleksitas komputasi dari algoritma RSC yang diusulkan adalah Θ(n2).

Algorithm 1. RSC (ring-stepping clustering) Algorithm

Inputs: (1) a set of GNSS points **PointSet** = {$P_1, P_2, P_3, \ldots, P_n$ }; (2) One parameter **R**.
Outputs: A set of centerlines **CenterLineSet** = {$CL_1, CL_2, CL_3, \ldots, CL_m$ }, and each centerline is composed of a set of points NodeSet = {$N_1, N_2, N_3, \ldots, N_k$ }

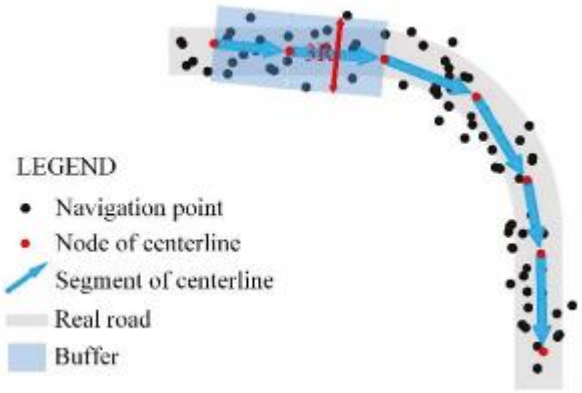
```
k–dtree ← Build k–d tree for PointSet
For i ← 1 to n do
  PointSet1 ← GetPointSetInRing(PointSet[i], 0, 2*R, PointSet)
  for j ← 1 to length(PointSet1) do
    PointSet1[j]'s PointSet2 = GetPointSetInRing(PointSet1[j], 0, R, PointSet)
    node ← the point whose PointSet2 has the most points in PointSet1
    numpts ← length of node's PointSet2
    If numpts > 0 then
      while numpts > 0 do
        add node to NodeSet
        PointSet3 ← GetPointSetInRing(Node, R, 3*R, PointSet)
        delete those points in PointSet3 that would make the centerline turn around
        for k ← 1 to length(PointSet3) do
          PointSet3[k]'s PointSet4 ← GetPointSetInRing(PointSet3[k], 0, R, PointSet3)
        end
        node ← the point whose PointSet4 has the most points in PointSet3
        numpts ← length of node's PointSet4
      end
    end
    add NodeSet to CenterLineSet
  end
end
return CenterLineSet
function GetPointSetInRing(Point, InsideRadius, OutsideRadius, PointSet)
  PointSetResult ← points in PointSet whose distance to Point is larger than InsideRadius and smaller than OutsideRadius (using kdtree)
return PointSetResult
```

Penghindaran dan Peningkatan Duplikasi

Dengan menggunakan algoritma yang diusulkan di atas, garis tengah yang berasal dari titik-titik GNSS diperoleh dan diatur berdasarkan urutan node. Namun, setelah ekstraksi garis

tengah jalan yang hilang tersebut, titik GNSS di jalan tersebut masih dicadangkan. Hal ini menyebabkan duplikasi garis tengah.

Oleh karena itu, buffer persegi panjang dengan lebar 3R untuk setiap segmen garis tengah dilakukan untuk menghilangkan titik-titik yang tertutup (lihat persegi panjang biru pada Gambar 8.4). Semua titik yang termasuk dalam buffer persegi panjang telah dihapus dari PointSet dan harus dihitung lebih lanjut. Selain itu, algoritma ekstraksi garis tengah dihentikan ketika PointSet kosong.



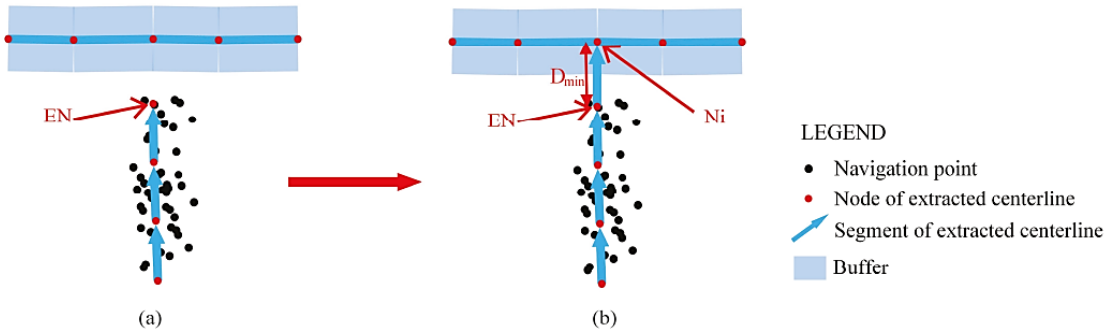
Gambar 8.4. Penghindaran dan peningkatan duplikasi garis tengah.

8.4 PEMBENTUKAN TOPOLOGI JALAN

Setelah mengekstraksi garis tengah jalan raya yang hilang, topologi jalan raya, yang penting untuk peta jalan, tetap hilang. Pada subbagian ini, topologi jalan dibangun kembali untuk jalan yang diekstraksi. Dua bagian utama dimasukkan: pembentukan topologi untuk jalan-jalan yang hilang yang diekstraksi dan pembentukan hubungan untuk jalan-jalan yang ada.

Pembentukan Topologi untuk Jalan yang Hilang

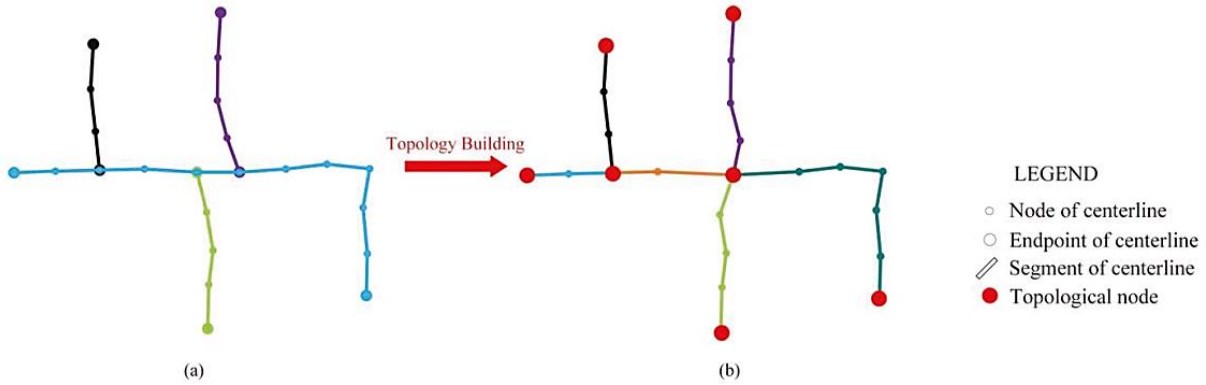
Membangun hubungan topologi dari garis tengah yang diekstraksi terdiri dari tiga langkah utama. Pertama, hubungan topologi antara garis tengah yang berbeda dibangun sesuai dengan simpul akhir (EN) dari garis tengah tersebut. Jarak antara node akhir (EN) dan node normal (Ni) dihitung, dan jarak minimum (D_{min}) ditentukan. Jika D_{min} < 3R maka hubungan antara EN dan Ni dianggap sebagai garis tengah, lihat Gambar 8.5.



Gambar 5. Hubungan yang dibangun antara garis tengah yang diekstraksi. (a) Sebelum membangun dan (b) setelah membangun. EN adalah node akhir, D_{min} adalah jarak minimum, Ni adalah mode normal.

Kemudian, analisis cluster digunakan untuk menggabungkan node topologi yang berdekatan sesuai dengan jarak antar node. Dalam tulisan ini, algoritma mean shift clustering diadopsi. Node topologi dalam jarak R diubah menjadi node topologi tunggal. Hal ini sangat meningkatkan hubungan topologi jalan yang hilang.

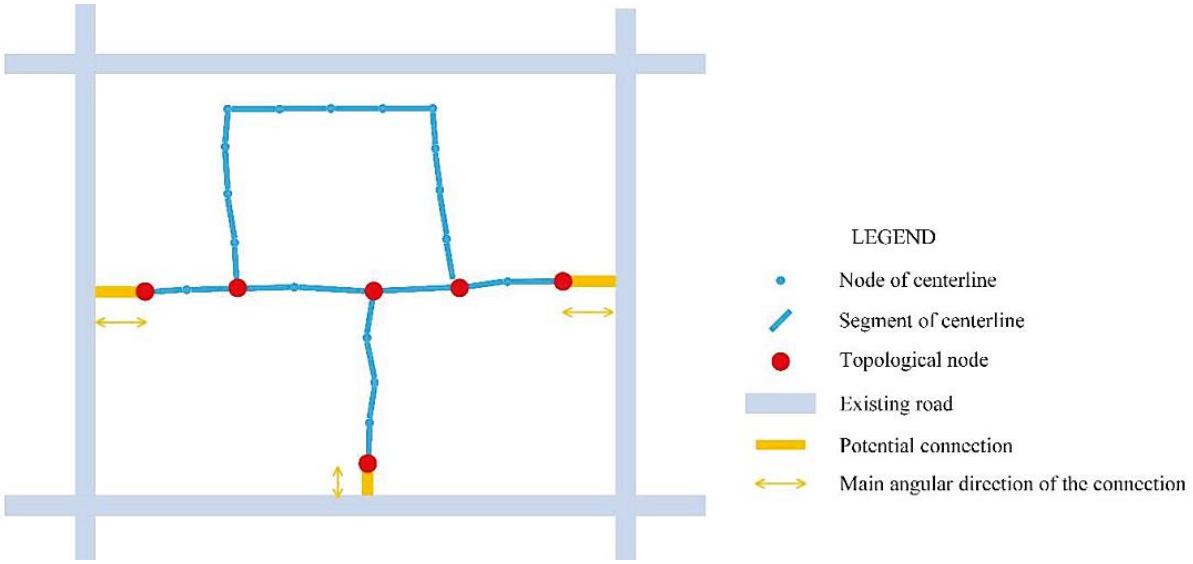
Terakhir, prosedur klasifikasi node dilakukan untuk mengklasifikasikan node menjadi dua kelompok: node topologi dan normal. Perbedaan antara node normal dan topologi adalah segmen garis tengah yang terhubung. Setelah segmen garis tengah yang terhubung dari sebuah simpul >2, simpul tersebut dianggap sebagai simpul topologi, dan garis tengah tersebut dibagi menjadi dua garis tengah anak (lihat simpul merah pada Gambar 8.6b).



Gambar 8.6. Pembangunan kembali topologi garis tengah yang diekstraksi. Garis tengah ditunjukkan (a) sebelum pembangunan kembali dan (b) setelah pembangunan kembali.

Hubungan antara Jalan yang Hilang dan Jalan yang Ada

Setelah topologi jalan yang hilang ditetapkan, hubungan antara jalan yang hilang dan jalan yang ada juga dibuat. Pertama, koneksi potensial dibangun antara titik-titik tersebut dan jalan yang ada (lihat garis kuning pada Gambar 8.7). Kemudian dilakukan verifikasi sambungan potensial dengan menggunakan distribusi azimuth titik-titik GNSS di sekitar sambungan potensial; setelah parameter ini seragam dengan arah sudut utama sambungan, sambungan potensial dianggap valid. Jika tidak, koneksi tersebut dianggap sebagai koneksi yang tidak valid.



Gambar 8.7. Contoh ilustrasi hubungan antara jalan yang hilang dan jalan yang ada.

8.5 DATA NAVIGASI SELULER

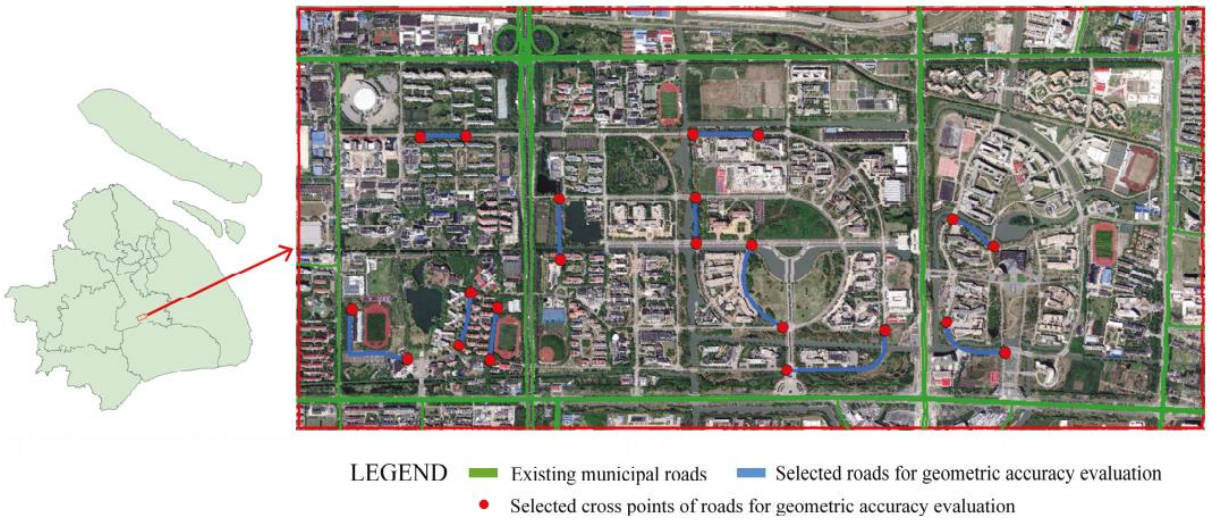
Dalam penelitian ini data navigasi seluler digunakan untuk mengekstraksi jalan yang hilang. Data ini dihasilkan dan dikumpulkan melalui telepon seluler. Data ini dihasilkan ketika aplikasi navigasi di ponsel terbuka, terlepas dari apakah pengguna menggunakannya untuk navigasi. Kecepatan pengambilan sampel adalah satu detik per catatan, dan setiap catatan berisi tujuh bidang utama: hari, waktu, ID, garis bujur, garis lintang, kecepatan, dan azimuth. Deskripsi masing-masing bidang diberikan pada Tabel 8.1.

Tabel 8.1. Deskripsi kolom data.

Indeks	Nama Bidang	Keterangan
1	Hari	Hari pembuatan catatan, termasuk tahun, bulan, dan hari
2	Waktu	Waktu rekaman dibuat, termasuk jam, menit, dan detik
3	Pengenal	ID pengguna
4	Garis bujur	Bujur posisi pengguna
5	Garis Lintang	Lintang posisi pengguna
6	Kecepatan	Kecepatan pengguna
7	Azimut	Azimuth pengguna

Kumpulan Data Kasus

Sebuah wilayah yang terletak di Shanghai dipilih sebagai wilayah kasus. Panjangnya 3,6 km dan lebar 1,6 km, serta meliputi area seluas 5,76 kilometer persegi. Area kasus mencakup dua kampus universitas, dan sebagian besar jalan di area kasus merupakan jalan dua arah dua jalur. Di sekitar area kasus, tersedia beberapa jalan kota, seperti Jalan Tol Hujin, Jalan Jianchuan, Jalan Dongchuan, dan Jalan Selatan Lianhua. Lokasi, citra, dan jalan kota yang ada (ditandai dengan warna hijau) di wilayah kasus ditunjukkan pada Gambar 8.8.



Gambar 8.8. Lokasi, citra, jalan kota yang ada di wilayah kasus, dan jalan serta titik persimpangan yang dipilih untuk evaluasi kualitas. Citra tersebut dikumpulkan pada 13 April 2017 oleh GeoEye-1 dari DigitalGlobe. Resolusi spasial citra adalah 1 m. Citra tersebut dicocokkan dengan peta sebenarnya dengan toolbar georeferensi perangkat lunak ArcGIS. Dibandingkan dengan peta sebenarnya, 13 titik kontrol pada gambar memiliki akurasi posisi 1,97 m RMSE (besar kesalahan berkisar antara 1,35 hingga 2,93 m).

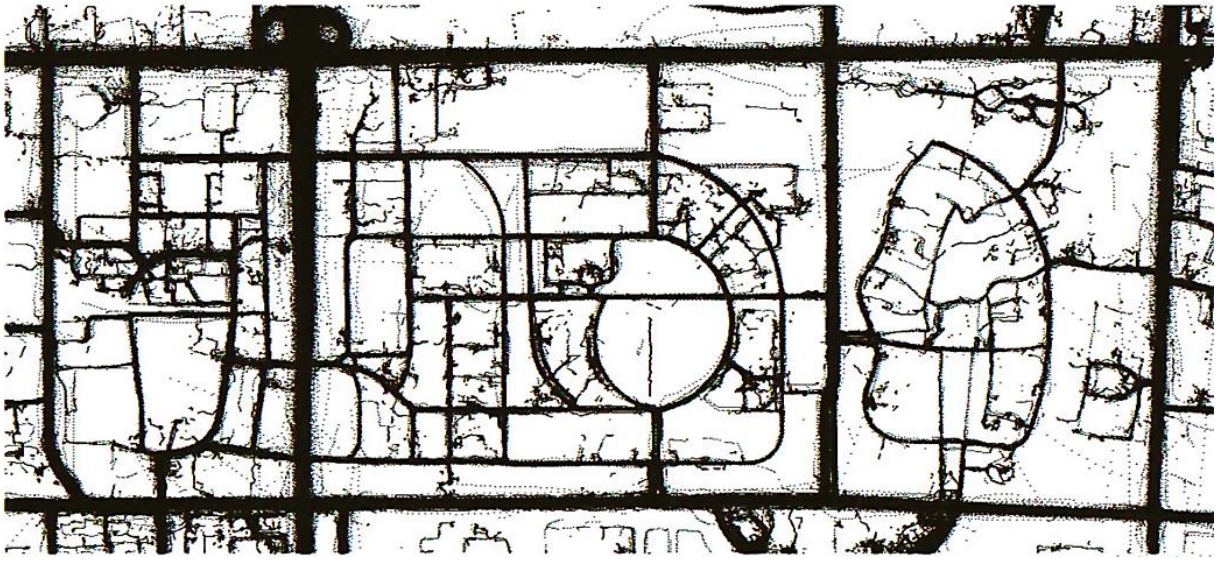
Untuk mengevaluasi kualitas jaringan jalan yang diekstraksi, digunakan peta wilayah sebenarnya yang disediakan oleh Institut Survei dan Pemetaan Kota Shanghai. Peta diukur dengan total stasiun secara manual, dan presisi peta mencapai tingkat sentimeter. Di antara semua jalan sebenarnya, 11 jalan, termasuk jalan lurus dan melengkung, dan 22 titik persimpangan (titik akhir dari 11 jalan yang dipilih) dipilih untuk mengevaluasi kinerja metode yang diusulkan. Untuk membandingkan kualitas metode yang diturunkan dari satelit, 11 jalan terpilih dipetakan secara manual dari citra satelit menggunakan perangkat lunak ArcGIS—pemetaan tersebut dilakukan oleh tiga operator yang masing-masing memiliki pelatihan yang baik dalam ekstraksi objek citra penginderaan jauh. Jalan yang dipilih dan titik persimpangan ditunjukkan pada Gambar 8.8. Jalan yang dipetakan secara manual dibandingkan dengan peta wilayah sebenarnya untuk mengevaluasi keakuratan spasial, yang ditunjukkan pada Tabel 8.2. Berdasarkan hasil evaluasi, ketepatan posisi jalan digitalisasi sekitar 1,00 m. Dibandingkan dengan data jalan sebenarnya, perbedaan sebenarnya adalah sekitar 3,95 m.

Tabel 8.2. Hasil akurasi spasial 22 titik jalan yang dipetakan secara manual. Kebenaran adalah jarak antara posisi rata-rata dan posisi sebenarnya yang bersangkutan (m). Presisi adalah kesalahan kuadrat rata-rata dari titik-titik oleh operator yang berbeda (m).

Point ID	Trueness (m)	Presisi (m)	Point ID	Trueness (m)	Presisi (m)
1	4.19	0,64	**12**	6.23	1.13
2	3.26	1.04	**13**	5.18	1.91
3	1.15	1.43	**14**	0,70	1.65
4	6.84	1.52	**15**	3.63	0,97
5	5.88	0,84	**16**	2.21	1.77
6	6.00	1.12	**17**	2.78	1.08

7	5.78	0,19	**18**	5.01	0,04
8	2.76	0,35	**19**	3.41	0,06
9	3.51	0,05	**20**	2.51	1.70
10	5.35	1.04	**21**	3.25	0,23
11	2.95	1.86	**22**	4.27	1.43
Rata-Rata	3.95	1,00			

Data yang digunakan untuk analisis dikumpulkan pada bulan Desember 2017 (11-15 Desember) dan terdiri dari total 9.944.710 titik data GNSS, yang mencakup 198.241 ID kendaraan unik. Data disediakan oleh 1RenData (ShangHai) Technology Co., Ltd (Shanghai, Cina). Gambar 9 menunjukkan distribusi data mentah GNSS.



Gambar 8.9. Data penelitian mentah. Titik hitam melambangkan titik GNSS (Global Navigation Satellite System).

Penyaringan Data

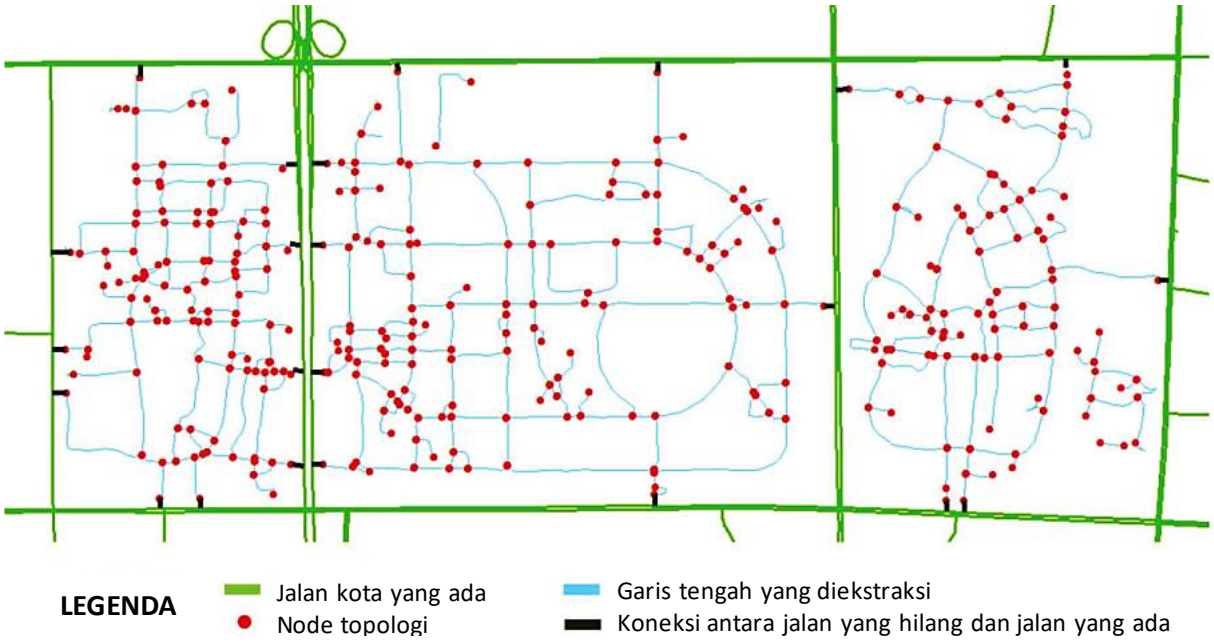
Dengan menggunakan metode pemfilteran data yang dijelaskan di Bagian 3.2, 392.128 catatan milik 8676 ID pengguna unik akhirnya dicadangkan. Poin GNSS yang dicadangkan mengambil ~4,0% dari data mentah. Komponen data kasus mentah ditunjukkan pada Tabel 8.3.

Tabel 8.3. Komponen data kasus.

Lokasi		Pejalan kaki	Kendaraan	Total
Jalan kota yang sudah ada	Nilai	3.054.797	6.179.641	9.234.438
	Presentase	30,7%	62,1%	92,8%
Jalan yang hilang	Nilai	318.144	392.128	710.272
	Presentase	3,2%	4,0%	7,2%
Total	Nilai	3.372.941	6.571.769	9.944.710
	Presentase	33,9%	62,1%	100,0%

8.6 EKSTRAKSI GARIS JALAN

Metode RSC yang diusulkan digunakan untuk mengekstraksi garis tengah jalan dari data setelah penyaringan. Prototipe metode diimplementasikan dalam bahasa pemrograman C-Sharp. Percobaan dilakukan pada server dengan CPU Intel Xeon Platinum 8163@ 2,5 GHz dan memori 16 GB. Karena lebar rata-rata jalan di area percobaan adalah ~7 m, parameter R ditetapkan ke 7,0 dalam bab ini. Garis tengah yang diperoleh ditunjukkan pada Gambar 10. Jumlah garis tengah sebanyak 603 buah, dan rata-rata panjang garis tengah tersebut adalah 116,09 m. Diperlukan waktu 557 detik untuk mengekstrak garis tengah dari 392.128 titik.



Gambar 8.10. Hasil studi kasus.

Evaluasi Kualitas Geometris Garis Tengah Jalan

Evaluasi Geometris Garis Tengah yang Dihasilkan Dibandingkan dengan Peta Nyata. Pada subbagian ini, dua indeks dipilih sebagai ukuran kualitas jalan yang diekstraksi. Indeks pertama adalah panjang jalan. Asumsikan Lr adalah panjang jalan sebenarnya dan Lg adalah panjang jalan yang digunakan untuk evaluasi kualitas. Kemudian, kesalahan absolut antara panjang jalan yang berbeda dihitung dengan Persamaan (2):

$$\mathrm{Eg} = |\mathrm{Lr} - \mathrm{Lg}|$$

dimana Eg adalah kesalahan mutlak antara Lr dan Lg.

Indeks kedua adalah jarak antara masing-masing garis tengah jalan dan jalan sebenarnya yang bersangkutan. Pertama, luas wilayah antara jalan asli dan jalan yang diekstraksi dihitung (lihat area biru pada Gambar 8.11). Kemudian, nilai kualitas garis tengah Tg yang diekstraksi dihitung menggunakan Persamaan (3):

$$\mathrm{Tg} = \frac{\mathrm{Ag}}{\mathrm{Lr}}$$

dimana Ag adalah luas wilayah antara jalan yang dihasilkan dan jalan sebenarnya.

Gambar 8.11. Perhitungan jarak antara jalan nyata dan jalan yang diekstraksi.

Dengan menggunakan parameter pada Persamaan (2) dan (3), kami mencantumkan hasil evaluasi kualitas 11 jalan terpilih pada Tabel 8.4. Berdasarkan Tabel 8.4, rata-rata Eg adalah 1,84 m dan rata-rata Tg adalah 1,62 m, yang berarti menunjukkan bahwa metode yang diusulkan dapat mencapai hasil yang sangat baik.

Tabel 8.4. Hasil evaluasi geometri garis tengah yang dihasilkan (m).

ID jalan	Lr	Lg	Eg	TG
1	206,95	207,93	0,98	0,37
2	403,05	401,05	2,00	1,21
3	183,80	181,47	2,34	2,51
4	199,36	196,90	2,46	1,70
5	304,27	304,96	0,69	3,01
6	511,84	511,84	0,00	1,80
7	240,78	238,18	2,60	1,89
8	219,78	223,12	3,34	0,34
9	180,94	182,75	1,81	1,94
10	410,14	406,24	3,90	2,11
11	265,42	265,55	0,13	0,96
Rata-rata	284,21	283,64	1,84	1,62

Untuk membandingkan akurasi geometrik antara metode yang diusulkan dan digitalisasi dengan citra penginderaan jauh, digunakan perbedaan sebenarnya antara titik persimpangan jalan sebenarnya yang dipilih dan titik persimpangan jalan terkait, yang diberi nama Dg. Dalam tulisan ini, 22 titik persimpangan jalan dipilih untuk menghitung perbedaan sebenarnya dengan peta sebenarnya. Kemudian Dg minimum, maksimum, dan rata-rata dari kedua metode tercantum pada Tabel 8.5.

Tabel 8.5 Dg minimum, maksimum, dan rata-rata digitalisasi dan metode yang diusulkan (m).

Digitalisasi Dari Citra Penginderaan Jauh			**Metode yang di Usulkan**		
Minimum	Maksimum	Rata-rata	Minimum	Maksimum	Rata-rata
0,23	6,00	3,95	0,81	6,02	3,43

Berdasarkan Tabel 8.5, rata-rata Dg jalan yang dipetakan adalah 3,95 m, lebih besar dari metode yang diusulkan. Hal ini menunjukkan bahwa keakuratan metode yang diusulkan sedikit lebih baik dibandingkan metode digitalisasi.

Evaluasi F-Score

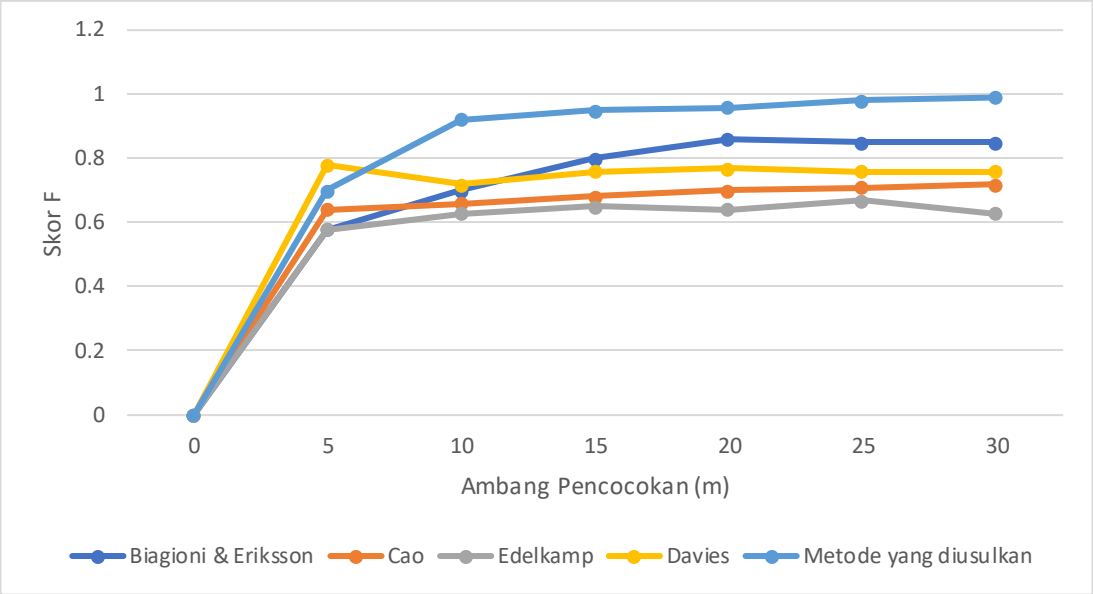
Selain metode evaluasi yang disebutkan di atas untuk garis tengah yang dihasilkan, skor F– yang diusulkan dalam juga diadopsi untuk mengevaluasi metode yang diusulkan. Skor F- dihitung sebagai berikut:

$$\text{spurious} = \frac{\text{spurious marbles}}{(\text{spurious marbles} + \text{matched marbles})}$$

$$\text{Missing} = \frac{\text{empty holes}}{(\text{empty holes} + \text{matched holes})}$$

$$\text{Skor F} = 2 \times \frac{(1 - \text{spurious})(1 - \text{missing})}{(1 - \text{spurious}) + (1 - \text{missing})}$$

Dimulai dari lokasi acak, jalan dieksplorasi dengan menempatkan titik sampel pada setiap grafik selama traversal keluar dalam radius maksimum. Titik sampel pada jalan yang memerlukan evaluasi dianggap sebagai “kelereng” dan pada jalan sebenarnya dianggap sebagai “lubang”. Dalam hal ini kelereng palsu melambangkan banyaknya titik pada jalan evaluasi yang tidak mendapat kecocokan, kelereng cocok melambangkan banyaknya titik pada jalan evaluasi yang mendapat kecocokan, lubang kosong melambangkan banyaknya titik pada jalan sebenarnya yang mendapat kecocokan. tidak mendapatkan kecocokan dan lubang yang cocok mewakili jumlah titik pada jalan yang dievaluasi yang mendapatkan kecocokan.



Gambar 8.12. F – skor metode yang diusulkan dan yang sudah ada.

Skor F- dari metode yang diusulkan dibandingkan dengan metode lainnya. Perbandingannya ditunjukkan pada Gambar 8.12. Jelas sekali, metode yang diusulkan menawarkan peningkatan yang signifikan dibandingkan metode sebelumnya.

Topologi Jalan Hilang

Topologi dibangun setelah ekstraksi garis tengah. Ada total 371 node topologi. Setelah verifikasi manual, 296 merupakan persimpangan jalan yang sebenarnya. Kebenarannya adalah ~79,8%. Node topologi yang diekstraksi secara salah, biasanya terletak di rumah-rumah, lokasinya terlalu padat atau terletak di jalan yang terlalu rumit. Hal ini mungkin disebabkan oleh rendahnya akurasi posisi dan efek multipath pada perangkat GNSS dan dapat ditingkatkan setelah menyempurnakan kumpulan data.

Hubungan antara Jalan yang Hilang dan Jalan yang Ada

Ditemukan dua puluh enam koneksi potensial antara node topologi dan jalan yang ada. Setelah verifikasi azimuth, 23 koneksi dicadangkan; jika dibandingkan dengan gambar penginderaan jauh yang terkait, gambar tersebut benar. Koneksi yang benar ini mewakili pintu masuk dan keluar.

8.7 RINGKASAN

Bab ini mengusulkan metode baru untuk menghasilkan jaringan jalan yang hilang di blok kota menggunakan data lintasan navigasi seluler yang besar. Algoritme bernama RSC dirancang berdasarkan data GNSS frekuensi tinggi. Setelah mengekstraksi garis tengah, diusulkan metode untuk membangun topologi jalan di blok kota dan membangun hubungan antara jalan yang hilang dan jalan yang ada. Area kasus (5,76 km2) digunakan untuk memverifikasi kelayakan dan validitas metode yang diusulkan. Hasilnya menunjukkan bahwa dibandingkan dengan jalan sebenarnya, perbedaan panjang rata-rata adalah sekitar 1,84 m dan jarak rata-rata adalah sekitar 1,64 m, menunjukkan bahwa metode yang diusulkan dapat mencapai hasil ekstraksi jalan yang hilang setinggi satu meter. Data dari jalan yang diekstraksi

dengan satelit menunjukkan bahwa metode yang diusulkan memberikan hasil yang lebih baik dibandingkan metode yang diperoleh dari citra. Sedangkan dengan menggunakan indeks F – score, metode yang diusulkan dapat mencapai hasil terbaik dibandingkan penelitian sebelumnya.

Kebaruan dari metode yang diusulkan adalah kualitas geometrik yang lebih tinggi dari jalan hilang yang diekstraksi. Selisih panjang dan jarak antara jalan yang diekstraksi dengan jalan sebenarnya masing-masing adalah sekitar 1,84 m dan 1,64 m. Hal ini memungkinkan kemungkinan terciptanya jaringan jalan yang rumit. Sementara itu, kinerja F – score dari metode yang diusulkan menunjukkan peningkatan yang besar dibandingkan metode sebelumnya, yang berarti bahwa jaringan jalan yang dihasilkan oleh metode yang diusulkan jauh lebih teliti.

Namun kompleksitas metode yang diusulkan adalah Θ(n2). Metrik ini menunjukkan bahwa ketika jumlah input data GNSS meningkat, konsumsi sumber daya dan waktu oleh algoritma juga akan meningkat secara geometris. Meskipun metode yang diperkenalkan dalam bab ini berhasil dengan baik di sebagian besar wilayah, hasil dan kualitasnya akan dipengaruhi oleh beberapa faktor.

Pertama, karena semrawutnya lintasan di tempat parkir, akan terjadi tumpukan titik GNSS di wilayah terkait. Oleh karena itu, jalan yang melalui tempat parkir mobil tidak dapat diekstraksi menggunakan metode yang diusulkan. Selain itu, keberadaan jembatan di atas jalan, tempat parkir bawah tanah, atau jalan bawah tanah di bawah jalan akan mengakibatkan kegagalan dalam menghasilkan garis tengah jalan di permukaan tanah.

Kedua, ketika jalan berada di antara dua gedung tinggi atau pepohonan, akan terbentuk kanopi yang sangat tebal, dan efek bayangan serta multijalur dapat menyebabkan koordinat titik-titik GNSS terbebani oleh kesalahan yang signifikan; dengan demikian, kualitas garis tengah yang diekstraksi melalui metode yang diusulkan akan buruk. Selain itu, ketika dua jalan berdekatan satu sama lain dan sejajar serta keduanya tidak cukup lebar, titik-titik GNSS yang terkait akan sulit dibedakan, sehingga mungkin akan diekstraksi sebagai satu jalan.

Terakhir, selama pemfilteran data, beberapa data GNSS dengan kecepatan <5,0 km/jam dihapus untuk memfilter pengguna pejalan kaki dan kendaraan. Pendekatan seperti ini akan mengakibatkan tidak disertakannya beberapa kendaraan yang melaju dengan kecepatan lebih rendah; dengan demikian, data GNSS yang terlibat dalam penghitungan juga akan berkurang.

BAB 9
BASIS PENGETAHUAN BERORIENTASI TUGAS PADA GEOGRAFIS

Dalam beberapa tahun terakhir, pesatnya perkembangan komputasi awan dan teknologi web telah membawa kemajuan yang signifikan pada rantai layanan informasi geografis (layanan GI) untuk memecahkan masalah geografis yang kompleks. Namun, pembangunan alur kerja pemecahan masalah memerlukan banyak keahlian bagi pengguna akhir. Saat ini, hanya sedikit penelitian yang merancang basis pengetahuan untuk menangkap dan berbagi pengetahuan pemecahan masalah geografis. Bab ini mengabstraksikan permasalahan geografis sebagai suatu tugas yang selanjutnya dapat diuraikan menjadi beberapa subtugas. Tugas ini membedakan tiga rincian berbeda: Geooperator, Tugas Atom, dan Tugas Komposit. Model tugas disajikan untuk menentukan garis besar solusi masalah pada tingkat konseptual yang mencerminkan proses pemecahan masalah. Basis pengetahuan berorientasi tugas yang memanfaatkan pendekatan berbasis ontologi dibangun untuk menangkap dan berbagi pengetahuan tugas. Basis pengetahuan ini memberikan potensi untuk menggunakan kembali pengetahuan tugas ketika dihadapkan pada masalah serupa. Secara meyakinkan, rincian implementasi dijelaskan melalui contoh analisis peringatan dini meteorologi.

9.1 PENDAHULUAN

Dalam beberapa tahun terakhir, dengan pesatnya perkembangan komputasi awan dan teknologi web, semakin banyak sumber informasi geografis (GIR), misalnya data geografis, fungsi analisis geografis, model, aplikasi, dll., yang telah dikemas ke dalam berbagai macam informasi. layanan informasi geografis (GIServices) yang dapat diakses oleh pengguna masyarakat umum melalui web. Misalnya, toolkit layanan web, bernama GeoPW, menyediakan serangkaian layanan geoproses, yang digunakan untuk memenuhi tugas pemrosesan data dan analisis spasial melalui infrastruktur informasi terdistribusi. Dalam komunitas geografis, Open Geospatial Consortium (OGC) menetapkan serangkaian spesifikasi antarmuka standar, seperti Web Feature Service (WFS), Web Map Service (WMS), Web Coverage Service (WCS), dan Web Processing Service (WPS), yang selanjutnya meningkatkan interoperabilitas dan berbagi GIServices berbasis web.

Dalam domain aplikasi geografis, permasalahan geografis biasanya berhubungan dengan data yang heterogen dan beberapa proses komputasi. Kemampuan GIService tunggal terbatas dan tidak dapat dilaksanakan secara efektif karena kompleksitas permasalahan geografis. Dalam dekade terakhir, pendekatan berbasis alur kerja telah berkembang menjadi cara utama untuk mengatasi permasalahan geografis yang kompleks. Saat ini, dengan bantuan spesifikasi antarmuka standar, GIServices yang diterbitkan oleh berbagai organisasi dapat dirangkai sebagai alur kerja geoproses yang dapat menggambarkan urutan pelaksanaan langkah-langkah pemecahan masalah dan meningkatkan kekuatan Atomic GIServices untuk memenuhi tugas-tugas geoproses yang kompleks. Secara umum, permasalahan geografis

memerlukan keahlian yang relatif mendalam, sehingga memerlukan para ahli untuk menyumbangkan pengetahuan pemecahan masalah mereka melalui alur kerja konseptual.

Dalam penelitian sebelumnya, sudah ada beberapa penyelidikan dalam formalisasi alur kerja dan interoperabilitas semantik untuk GIServices. Selain itu, sejumlah penelitian telah menggunakan konsep tugas untuk memfasilitasi ekspresi kebutuhan pengguna pada tingkat semantik. Faktanya, banyak permasalahan geografis yang memiliki alur kerja konseptual yang serupa. Oleh karena itu, alur kerja konseptual dapat diformalkan menjadi basis pengetahuan, yang dapat memfasilitasi pengguna di masa depan untuk memecahkan masalah serupa.

Dalam bab ini, kami fokus pada penggunaan ontologi yang dikaitkan dengan pendekatan berorientasi tugas untuk membangun basis pengetahuan guna meningkatkan pemecahan masalah geografis. Secara umum diyakini bahwa ontologi adalah fondasi dan bagian penting dari jaringan semantik. Ontologi menyediakan istilah terpadu untuk meningkatkan interoperabilitas semantik pengetahuan domain. Sebuah tugas diperkenalkan sebagai komponen yang dapat digunakan kembali untuk memodelkan urutan langkah-langkah inferensi yang terlibat dalam proses penyelesaian masalah geografis tertentu pada tingkat konseptual. Basis pengetahuan dapat menyimpan alur kerja konseptual yang dianggap sebagai pengetahuan apriori yang dikumpulkan dari pengalaman masa lalu para ahli domain, yang dapat memungkinkan pengetahuan pemecahan masalah dapat digunakan kembali. Masalah geografis diabstraksikan sebagai sebuah tugas, dan pengetahuan untuk tugas tersebut dianggap sebagai solusi masalah. Dalam banyak keadaan, tugas-tugas perlu didekomposisi menjadi tugas-tugas yang lebih sederhana, yang masing-masing dapat diselesaikan dengan satu atau serangkaian fungsi. Karena tugas yang lebih kecil lebih sederhana daripada tugas keseluruhan, maka kompleksitas tugas berkurang secara signifikan. Oleh karena itu, kami selanjutnya membagi tugas menjadi tiga rincian berbeda:

1. Geooperator, yang merupakan fungsi pemrosesan dasar;
2. Tugas atomik, yang tidak dapat diurai; dan
3. Tugas gabungan, yang dipecah menjadi beberapa subtugas.

Pekerjaan utama dari bab ini mencakup hal-hal berikut:

1. *Konsep:* konsep tugas diperkenalkan sebagai komponen yang dapat digunakan kembali untuk pemecahan masalah geografis dan digunakan untuk mencerminkan kebutuhan pengguna;
2. *Model:* model tugas diusulkan untuk mensimulasikan proses pemecahan masalah;
3. *Basis pengetahuan:* basis pengetahuan ontologis dirancang, yang terdiri dari beberapa ontologi yang dapat dioperasikan untuk menangkap dan berbagi pengetahuan pemecahan masalah; dan
4. *Implementasi:* dengan mengambil analisis peringatan dini meteorologi (MEW), kami menjelaskan rincian implementasi secara meyakinkan. Kami fokus pada solusi masalah geografis sebagai tugas yang terdiri dari operasi geoproses konseptual yang tidak berhubungan dengan layanan konkret apa pun. Instansiasi dan eksekusi tugas,

interaksi tingkat rendah dengan operasi (seperti mengakses data input), dan validasi rantai pemrosesan tidak termasuk dalam cakupan bab ini.

9.2 PENDEKATAN BERBASIS TUGAS

Gagasan tugas ini diusulkan oleh Albrecht di bidang sistem informasi geografis pada awal tahun 1990an dan telah digunakan dalam banyak penelitian. Namun, masih belum ada definisi terpadu tentang suatu tugas. Secara umum, konsep tugas mencerminkan kebutuhan pengguna dan menjelaskan semua tindakan atau operasi untuk memecahkan masalah tertentu. Beberapa penelitian telah dilakukan dengan menggunakan pendekatan berbasis tugas. Kami merangkum dan mengklasifikasikannya sebagai berikut:

1. *Bahasa berbasis tugas.* Bahasa ontologi tugas berdasarkan OWL (Web Ontology Language), bernama OWL-T, telah diusulkan untuk mendefinisikan templat tugas untuk memformalkan permintaan pengguna dan proses bisnis pada abstraksi tingkat tinggi, yang digunakan untuk tugas rencana perjalanan. Hu dkk. memperluas pendekatan berorientasi tugas ke domain Web Sensor OGC. Bahasa Model Tugas, yang disebut TaskML, adalah bahasa untuk tugas pemodelan. Fitur penting dari TaskML adalah Pemicu Tugas, Prioritas Tugas, dan QoS Tugas.
2. *Pendekatan ontologi tugas.* Matahari dkk. mengusulkan pendekatan berbasis ontologi tugas untuk domain geografis untuk mewujudkan geoproses langsung dalam lingkungan berorientasi layanan, yang mencakup tiga langkah: pembuatan model tugas, pembuatan model proses, dan eksekusi alur kerja. Sebuah studi kasus analisis banjir digunakan untuk menggambarkan efek dan peran tugas tersebut. Liu mengusulkan model ontologi tugas untuk manajemen dialog yang tidak bergantung pada domain dan menciptakan manajer dialog yang tidak bergantung pada tugas. Taman dkk. menyajikan ontologi tugas berdasarkan perspektif wisatawan menggunakan tugas, aktivitas, hubungan, dan properti. Sebuah sistem prototipe dikembangkan menggunakan menu berorientasi tugas.
3. *Pendekatan berbasis tugas untuk akuisisi data geografis.* Wiegand dan García mengusulkan pendekatan berbasis tugas untuk memajukan pengambilan sumber data geografis. Lebih konkretnya, mereka merancang model konseptual yang menggabungkan ontologi tugas, sumber data, metadata, dan tempat serta menggunakan mesin aturan Jess dan alat Protégé untuk menyediakan pemrosesan otomatis untuk pengambilan data. Qiu dkk. mengusulkan pendekatan berorientasi tugas untuk pengelolaan data bencana yang efisien yang melakukan pemetaan dari tugas darurat ke sumber data dan menghitung korelasi antara kumpulan data dan tugas umum. Contoh darurat banjir menggambarkan penggunaan pendekatan ini.

Pemecahan Masalah Geografis

Saat ini, teknologi layanan geoproses digunakan secara luas untuk memecahkan masalah geografis tertentu dalam infrastruktur informasi terdistribusi. Banyak penelitian telah dikhususkan untuk memanfaatkan atau memfasilitasi layanan geoproses untuk mendukung pemecahan masalah. Mikita menerbitkan layanan geoproses bagi pemilik hutan untuk

mengoptimalkan ukuran dan bentuk tebang habis selama proses pemulihan hutan. Müller mengusulkan kerangka hierarki untuk mengidentifikasi sifat semantik dan sintaksis layanan geoproses dengan empat tingkat granularitas, yang kondusif untuk pengambilan layanan, perbandingan layanan, dan pemanggilan layanan.

Dalam kebanyakan kasus, layanan geoproses tunggal tidak cukup untuk memecahkan masalah geografis yang kompleks. Oleh karena itu, teknologi alur kerja geoproses memberikan solusi. Integrasi layanan geoproses telah menjadi topik penelitian populer, dan serangkaian alat dan arsitektur dikembangkan untuk mendukung rangkaian layanan geoproses. Misalnya, alat alur kerja geoproses sumber terbuka, bernama GeoJModelBuilder, mampu mengintegrasikan layanan geoproses yang dapat dioperasikan dan menyusunnya menjadi alur kerja. Mesin orkestrasi RichWPS yang dikombinasikan dengan DSL (Domain Specific Language) digunakan untuk mengatur proses WPS dan mempublikasikan komposisi sebagai proses WPS untuk komposisi selanjutnya. Selain itu, ada banyak sistem manajemen alur kerja yang populer untuk memfasilitasi integrasi layanan geoproses, seperti Taverna, Triana, Kepler, jABC. Namun, mereka hanya menyederhanakan proses konstruksi alur kerja pada tingkat sintaksis, dan membangun alur kerja yang terdiri dari layanan untuk penyelesaian masalah geografis masih merupakan tantangan bagi pengguna akhir.

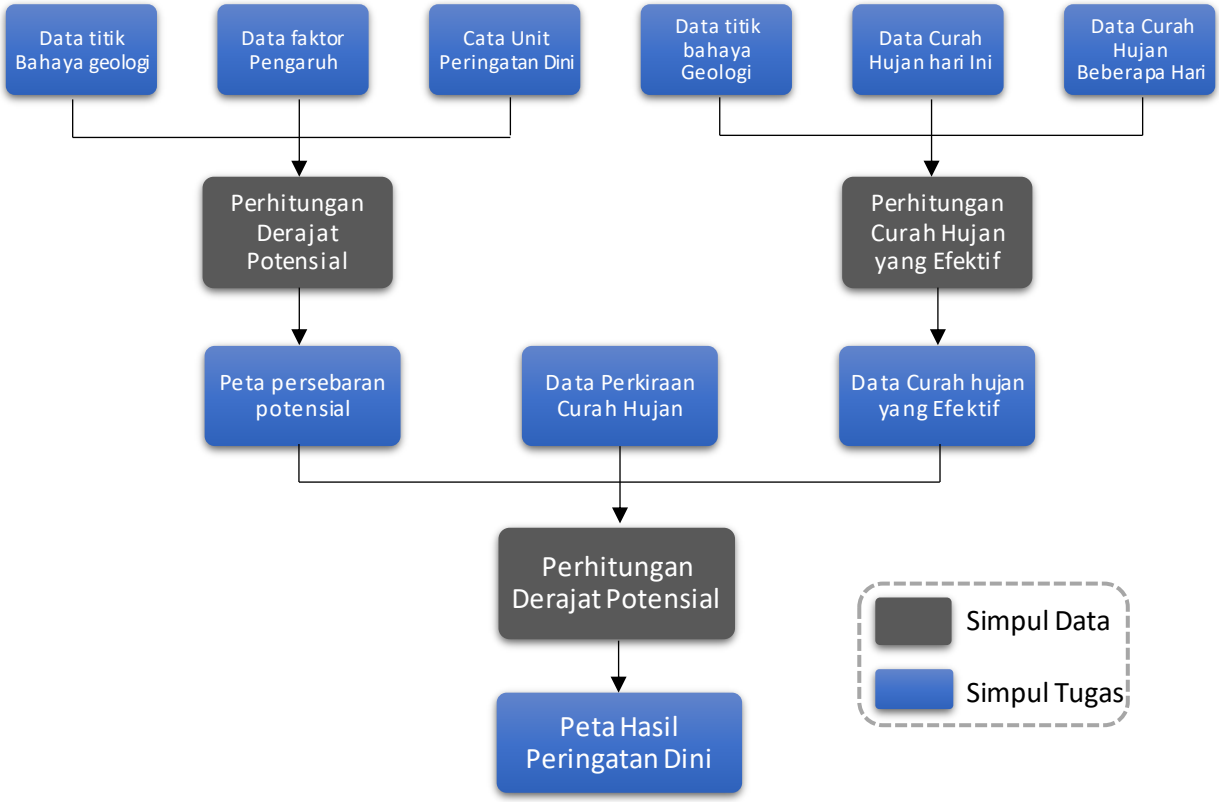
Baru-baru ini, lebih banyak penelitian yang berfokus pada komposisi alur kerja semantik dan otomatis untuk pemecahan masalah geografis. Farnaghi dan Mansourian mengusulkan solusi komposisi otomatis menggunakan algoritma perencanaan AI (Artificial Intelligence) dan SAWSDL (Semantic Annotations for Web Service Description Language) untuk meningkatkan proses manajemen bencana. Al-Areqi dkk. menerapkan metode sintesis berbasis kendala untuk menerapkan komposisi alur kerja semi-otomatis untuk analisis dampak kenaikan permukaan laut. Samadzadegan dkk. merancang kerangka kerja untuk alur kerja otomatis untuk deteksi dini kebakaran berdasarkan layanan OGC. Arul dan Prakash menyajikan algoritma komposisi terpadu yang menambahkan fase baru yang disebut Validasi dan Optimasi pada komposisi layanan web otomatis dan menghasilkan proses komposisi yang dapat diskalakan sesuai dengan perubahan dinamis kebutuhan pengguna.

9.3 SKENARIO APLIKASI

Di bagian ini, kami mendemonstrasikan contoh yang menggunakan alur kerja yang terdiri dari data terdistribusi dan berbagai layanan geoproses. Contoh ini digunakan sepanjang sisa bab ini untuk membantu memahami konsep tugas geografis. Dengan asumsi pengguna akhir adalah staf departemen pemantauan bencana meteorologi, maka ia perlu memprediksi kemungkinan terjadinya bencana geologi di suatu wilayah tertentu pada hari berikutnya. Hasil yang ideal adalah peta tematik wilayah peringatan dini yang menggunakan warna berbeda untuk mewakili tingkat peringatan dini yang berbeda.

Untuk mencapai hasil peringatan dini, pendekatan yang paling umum adalah dengan merumuskan alur kerja pemrosesan geografis yang dapat menghasilkan peta hasil peringatan dini. Seperti yang ditunjukkan pada Gambar 9.1, bentuk elips mewakili node data, dan bentuk persegi panjang mewakili node pemrosesan data. Pertama, menggunakan data titik bahaya

geologi, data faktor pengaruh, dan data satuan peringatan dini sebagai data masukan untuk menghitung indeks derajat potensi masing-masing unit peringatan dini. Demikian pula dapat memperoleh data curah hujan yang efektif. Kemudian peta sebaran potensi dan data curah hujan efektif dari langkah sebelumnya dengan data prakiraan curah hujan melalui perhitungan analisis peringatan dini untuk mencapai peta hasil peringatan dini.



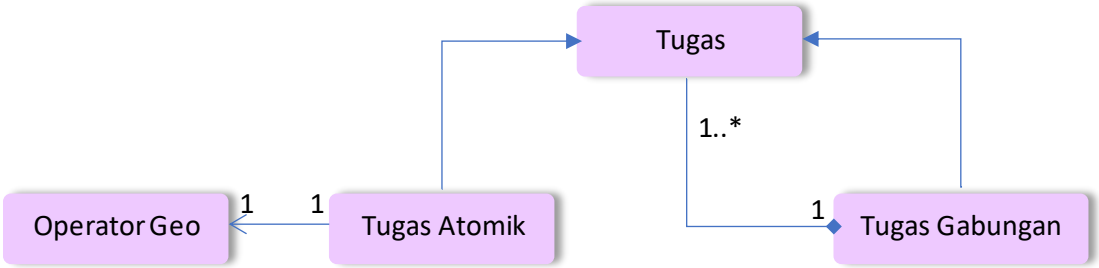
Gambar 9.1. Urutan proses peringatan dini meteorologi (MEW).

Untuk aplikasi yang disebutkan di atas, seluruh alur kerja dapat dianggap sebagai tugas. Pakar domain GIS dengan pengetahuan profesional mampu menganalisis prosedur teknologi dan mengabstraksikannya dalam bentuk konseptualisasi, yang kemudian digunakan untuk menggambarkan kerangka pengetahuan dari proses pemecahan masalah. Tugas MEW, yang sebelumnya dilakukan secara manual dan memerlukan keterampilan GIS serta pengetahuan proses bisnis, kini dapat dijalankan secara otomatis.

Tugas dan Model Tugas

Konsep tugas diusulkan untuk mencerminkan kebutuhan pengguna, yang dapat diselesaikan oleh satu atau lebih layanan geoproses. Masalah geografis disarikan sebagai tugas yang menunjukkan tujuan bisnis tingkat tinggi, dan pengguna menjalankan serangkaian proses untuk mencapai tujuan tersebut. Tugas berbeda dari operasi atau layanan, karena tugas berfokus pada apa yang ingin diselesaikan pengguna, sedangkan operasi atau layanan terutama berfokus pada penerapan komputasi geoproses.

Masalah yang kompleks dapat terdiri dari beberapa proses pemecahan masalah dengan persyaratan berbeda, sehingga sulit untuk mendefinisikan solusi sebagai satu tugas. Oleh karena itu, tugas yang kompleks dapat didekomposisi menjadi beberapa tugas yang lebih kecil, yang masing-masing dapat diselesaikan dengan cara yang relatif independen oleh satu atau lebih layanan geoproses dan kemudian digabungkan menjadi solusi yang lengkap. Perincian tugas memainkan peranan penting selama proses pemecahan masalah. Seperti yang ditunjukkan pada Gambar 9.2, ada tiga perincian yang berbeda: (1) operator geo sebagai fungsi dasar untuk tugas atom, (2) tugas atom sebagai blok penyusun untuk tugas gabungan, dan (3) tugas gabungan sebagai bangunan blok untuk tugas geografis yang kompleks. Konsekuensinya, tugas tersebut merupakan komponen yang dapat digunakan kembali untuk menyusun alur kerja pemecahan masalah.



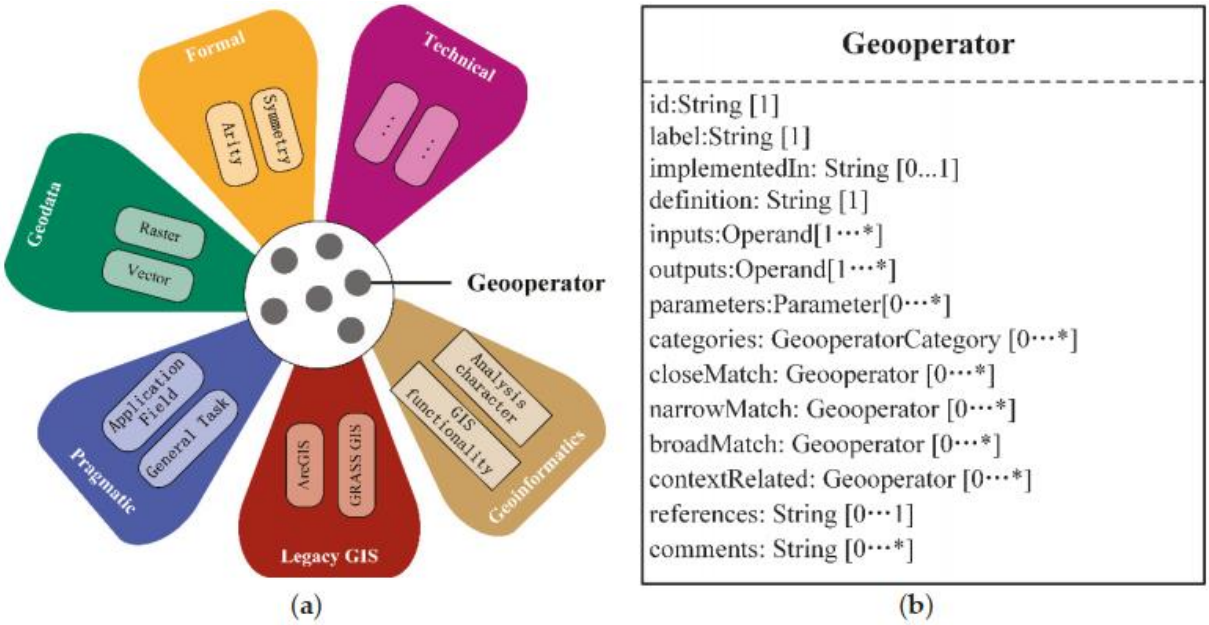
Gambar 9.2. Hubungan antara Task, AtomicTask, CompositeTask, dan Geooperator.

Properti proses dari tugas geografis dinyatakan dengan grafik proses tugas (TPG), yang digunakan untuk menangkap urutan pelaksanaan langkah-langkah penyelesaian masalah dan menggambarkan secara dekat bagaimana suatu tugas harus dicapai. Setiap TPG berisi sekumpulan sisi yang membentuk struktur grafik berarah asiklik. Tepi menunjukkan alur kerja dari dua tugas. Arah tepi menentukan ketergantungan antar tugas. Kombinasi TPG dan tugas membentuk model tugas yang menyediakan pendekatan yang memungkinkan pengguna menentukan masalah geografis yang kompleks pada tingkat abstrak.

Operator geografis

Pengetahuan pemecahan masalah geografis direpresentasikan pada tingkat konseptual yang memerlukan kategorisasi dan formalisasi layanan geoproses. Geooperator sebagian besar dikembangkan untuk meningkatkan kemudahan penemuan dan pertukaran fungsi geoproses dan menyediakan pendekatan untuk memformalkan fungsi geoproses yang terdefinisi dengan baik. Dalam karya Brauner, geooperator dikategorikan berdasarkan berbagai perspektif berbeda, seperti geodata, GIS warisan, perspektif pragmatis, formal, atau teknis. Ikhtisar perspektif dan kategori tingkat atas yang diidentifikasi oleh Brauner ditunjukkan pada Gambar 9.3a, dan elemen yang dijelaskan oleh geooperator diberikan pada Gambar 9.3b, yang dapat memfasilitasi pekerjaan kami. Yang pertama digunakan untuk mendefinisikan subkelas kelas Geooperator dalam ontologi operasi GIS tanpa modifikasi lebih lanjut; yang terakhir sebagian diubah menjadi properti data dan properti objek kelas Geooperator.

Geooperator diperkenalkan untuk memberikan konseptualisasi layanan geoproses (seperti analisis geografis atau layanan transformasi) yang dienkapsulasi sebagai layanan web standar (misalnya WPS) untuk menyediakan fungsionalitas geoproses di web. Dari perspektif berorientasi objek, operator geo bertindak sebagai pembungkus untuk layanan geoproses yang ada dan selanjutnya berfungsi sebagai blok penyusun untuk tugas-tugas geoproses dasar.



Gambar 9.3. (a) Perbedaan perspektif tentang Geooperator (b) Deskripsi elemen Geooperator.

9.4 DEFINISI FORMAL

Definisi 1 (Tugas). Sebuah tugas dapat didefinisikan sebagai empat kali lipat: (Persamaan 1)

$$\mathrm{T} = (\mathrm{PT}, \mathrm{OP}, \mathrm{PA}, \mathrm{C})$$

dimana PT menentukan jenis tugas, OP adalah input dan output spasial (misalnya, kumpulan data spasial), PA adalah sekumpulan parameter non-spasial dari suatu tugas, dan C terdiri dari prasyarat dan hasil yang secara umum membatasi atribut tematik dan geometris dari tugas tersebut. data input atau output untuk tugas geoproses.

Definisi 2 (Grafik Proses Tugas). Grafik proses tugas mendefinisikan struktur dasar dekomposisi tugas, yang merupakan grafik berarah asiklik yang didefinisikan sebagai berikut: (Persamaan 2)

$$\mathrm{TPG} = (\mathrm{V}, \mathrm{E})$$

dimana V adalah himpunan berhingga yang terdiri dari n simpul $\{ v_1, v_2, v_3 \ldots v_n \}$, dan setiap simpul $v \in \mathrm{V}$ menyatakan suatu tugas t_v. E adalah himpunan berhingga dari sisi berarah $\{e_{vi,vj}\}$. Setiap sisi $e_{vi,vj} \in E$ dapat dikarakterisasi dengan tuple $(p_{vi,vj}, c_{ij}) \cdot P_{vi,vj} = \langle v_i, v_j \rangle$ adalah pasangan terurut yang mewakili prioritas eksekusi antara tugas t_{vi} dan tugas t_{vj};

dengan kata lain, t_{vi} berada di depan t_{vj} pada urutan penguraian tugas yang juga dapat dilambangkan dengan $v_i < v_j$. c_{ij} mewakili penghubung aliran kontrol antara dua tugas, yang mencakup urutan, percabangan, loop, dan sebagainya.

Definisi 3 (Model Tugas). Model tugas didefinisikan oleh 2 tupel sebagai berikut (Persamaan 3)

$$\mathrm{TModel} = (\mathrm{t}, \mathrm{tpg})$$

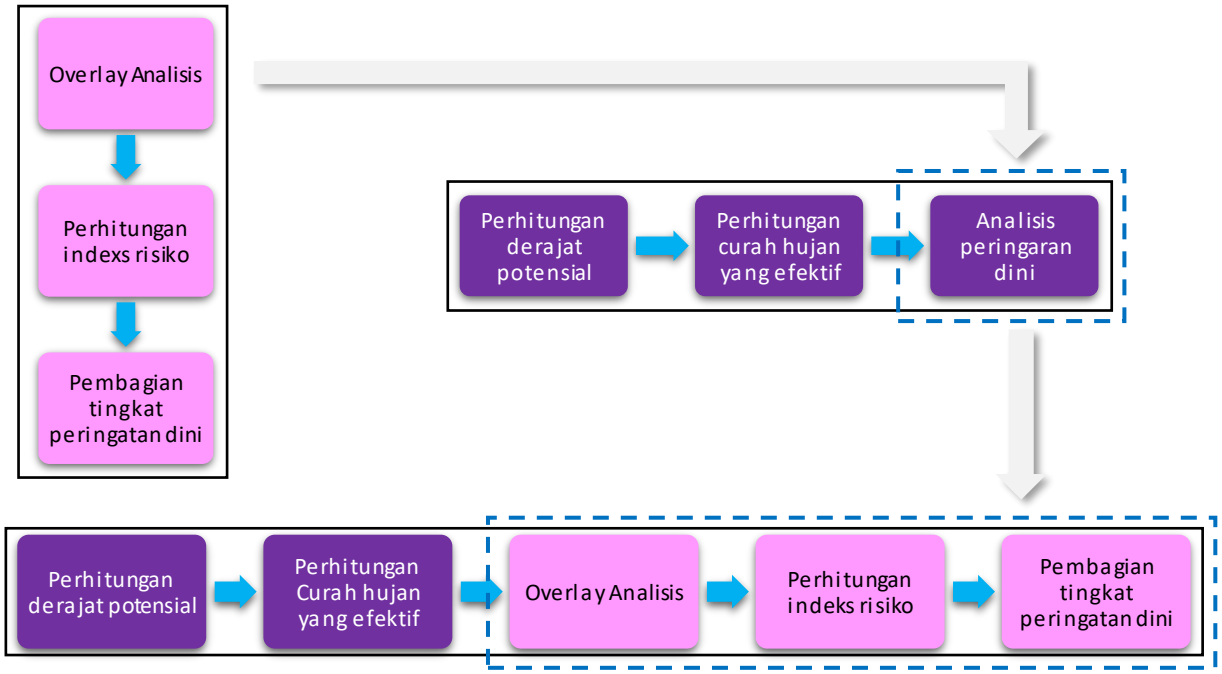
dimana $t \in \mathrm{T}$ adalah contoh tugas, dan tpg menunjukkan grafik proses tugas yang terkait dengan t yang mendefinisikan struktur dekomposisi. Jika tpg hanya berisi geooperator, kami menganggap tugas ini sebagai tugas atom; jika tidak, ini adalah tugas gabungan.

Definisi 4 (Dekomposisi Tugas). Mengikuti definisi model tugas, kita selanjutnya dapat menyelesaikan dekomposisi tugas. Diberikan grafik proses tugas $\mathrm{tpg} = V, E$, dengan asumsi $v \in \mathrm{V}, t_n \in \mathrm{T}$, v berhubungan dengan t_v. Jika t_v memiliki model yang sesuai $tmodel_v = (t_v, tpg_v$ maka penguraian tugas dapat ditentukan dengan: (Persamaan 4)

$$\mathrm{tpg}' = \mathrm{Dekomposisi}(\mathrm{v}, \mathrm{tpg}, \mathrm{tmodel_v})$$

dimana tpg' adalah grafik proses tugas baru yang diperoleh dengan menggantikan node v dengan $\mathrm{tpg_v}$ di $\mathrm{tmodel_v}$.

Mengambil alur kerja yang disebutkan sebagai contoh, Gambar 9.4 menggambarkan prosedur dekomposisi tugas. Node "Analisis peringatan dini" digantikan oleh grafik proses tugas, yang didefinisikan dalam model tugas, dimana sisi yang sebelumnya terhubung dengan node ini direvisi.



Gambar 9.4. Contoh dekomposisi tugas.

9.5 BASIS PENGETAHUAN BERORIENTASI TUGAS

Bagian ini menyajikan basis pengetahuan yang mengadaptasi pendekatan berbasis ontologi dan memberikan pengetahuan komprehensif untuk mendukung tugas geoproses. Untuk membangun basis pengetahuan, diperlukan seperangkat ontologi untuk menangkap pengetahuan yang terkait dengan solusi pemecahan masalah. Penggunaan ontologi membuat makna semantik dari prosedur penyelesaian masalah menjadi eksplisit dan lebih memudahkan pengguna untuk mendapatkan solusi masalah. Memformalkan basis pengetahuan akan membantu pengguna non-spesialis GIS dan spesialis dalam mengotomatisasi pemecahan masalah, memungkinkan penggunaan kembali dan berbagi solusi. Oleh karena itu, kami menganggap bahwa basis pengetahuan itu berharga.

Latar Belakang Ontologi

Telah diketahui secara luas bahwa ontologi menyediakan bahasa formal untuk membakukan dan berbagi semantik berbagai jenis pengetahuan domain. Kata ontologi pertama kali digunakan sebagai konsep filosofis dan membahas hakikat keberadaan, dan kemudian diperkenalkan ke dalam domain informasi oleh para peneliti. Saat ini, salah satu definisi ontologi yang paling umum adalah "Ontologi adalah spesifikasi eksplisit dari suatu konseptualisasi", yang diusulkan oleh Gruber pada tahun 1993. Berdasarkan definisi ini, ontologi pada dasarnya adalah taksonomi dunia objektif dan model representasi pengetahuan. Sementara itu, ontologi juga mendukung hubungan non-taksonomi.

Menurut Perez, pengetahuan dalam ontologi diformalkan oleh lima jenis pemodelan primitif: konsep, relasi, fungsi, aksioma, dan contoh. Dari sudut pandang matematis, ontologi secara formal dapat dinyatakan dengan Persamaan sebagai berikut: (Persamaan 5)

$$\mathrm{O} = \{\mathrm{C}, \mathrm{R}, \mathrm{F.A.I}\}$$

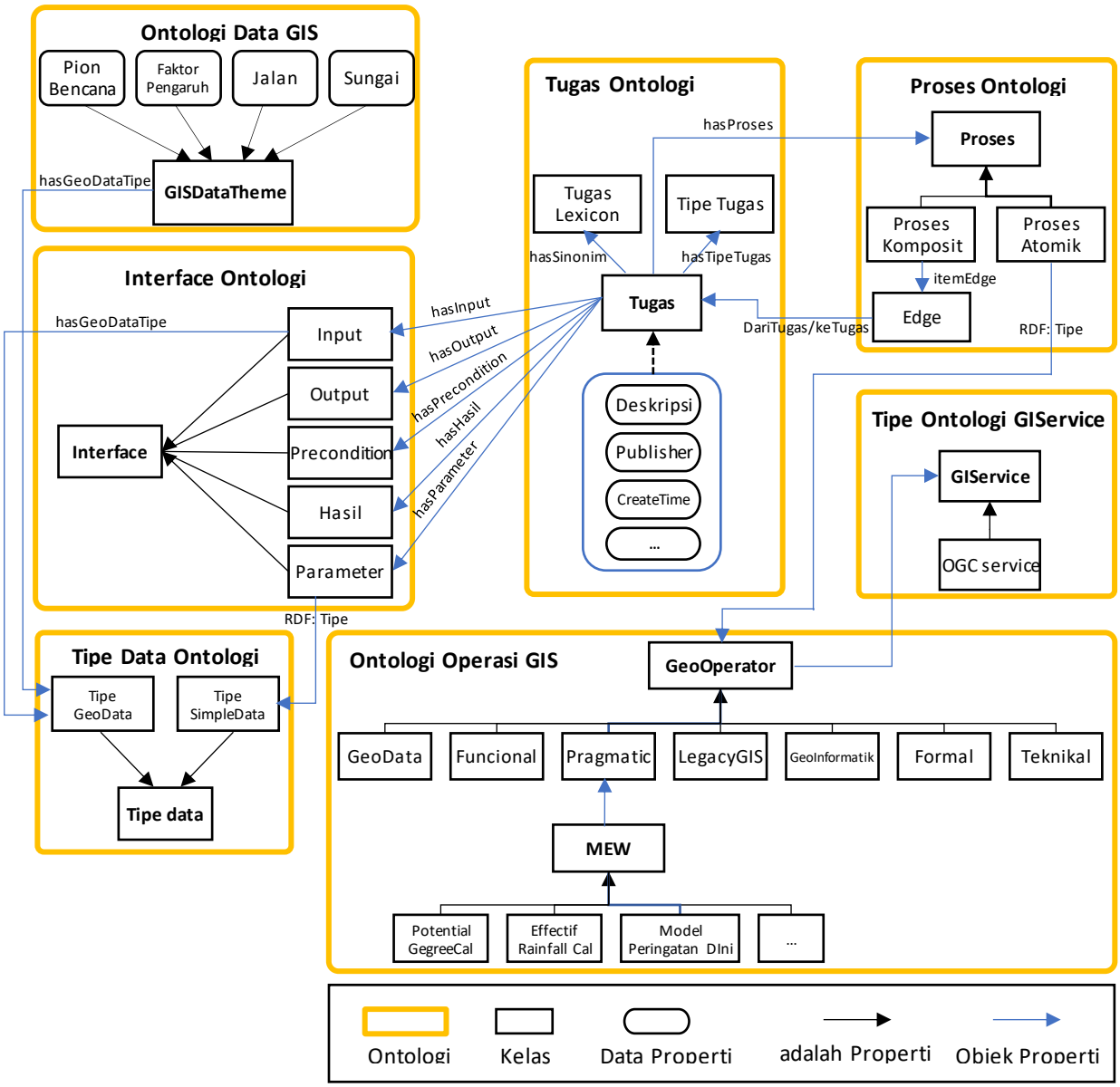
dimana C adalah himpunan yang unsur-unsurnya disebut konsep; R adalah himpunan hubungan antar konsep, $\mathrm{R} \subseteq \mathrm{C} \times \mathrm{C}$; F adalah relasi istimewa yang elemen-elemen sebelumnya n − 1 dapat menentukan elemen ke-n secara unik, dan dapat didefinisikan sebagai berikut: $\mathrm{F}\colon \mathrm{C}_1 \times \mathrm{C}_2 \times \ldots \times \mathrm{C}_{\mathrm{n}-1} \rightarrow \mathrm{C}_\mathrm{n}\colon \mathrm{A}$ mewakili aksioma geografis, yaitu kumpulan pernyataan dalam bentuk logis yang selalu benar; dan saya mewakili contoh konsep.

Dalam proses membangun ontologi, instance merepresentasikan objek yang bisa berupa apa saja dalam suatu domain, dan konsep adalah sekumpulan objek yang dipetakan ke kelas. Hubungan antar konsep diwujudkan oleh properti yang diklasifikasikan menjadi dua jenis: properti objek dan properti data. Properti objek menentukan hubungan antara dua kelas, dan menghubungkan dua individu dari kelas yang berbeda. Properti data mendefinisikan hubungan antara individu dan nilai data, yang mirip dengan atribut yang melekat pada suatu objek.

Ontologi di Inti Basis Pengetahuan

Untuk mewujudkan kemampuan merepresentasikan pengetahuan proses pemecahan masalah, basis pengetahuan menyediakan sekumpulan ontologi sebagai berikut: Ontologi Tugas, Ontologi Proses, Ontologi Operasi GIS, Ontologi Antarmuka, Ontologi Tipe Data, Ontologi Data GIS, dan Tipe GIService Ontologi. Ontologi-ontologi ini digabungkan untuk

memberikan dukungan bagi semua aspek pemecahan masalah, yang masing-masing memainkan peran kunci dalam membangun basis pengetahuan berorientasi tugas yang kaya, dinamis, dan fleksibel. Gambar 5 menunjukkan penggambaran definisi ontologi dan bagaimana mereka berhubungan satu sama lain. Beberapa ontologi penting dibahas secara rinci di bagian berikut.



Gambar 9.5. Hubungan ontologi dalam basis pengetahuan.

Ontologi Tugas

Ontologi Tugas adalah inti untuk mendukung pemecahan masalah, yang mendefinisikan kelas Tugas untuk mewakili masalah geografis. Relasi propertinya terdiri dari properti objek dan properti data. Properti data terutama menjelaskan informasi metadata

instans tugas, seperti Deskripsi, Penerbit, Waktu Pembuatan, dan seterusnya. Properti objek meliputi: hasSynonym, hasTaskType, hasProcess, hasInput, hasOutput, dll.

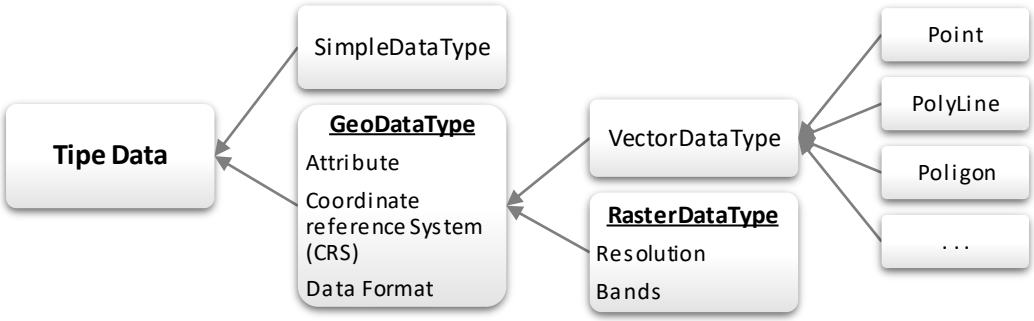
Kelas Task mengacu pada kelas Task Lexicon melalui properti hasSynonym untuk anotasi semantik tugas untuk menyediakan kata dan frasa yang mendeskripsikan tugas, yang menjadi dasar pengguna akhir mengeksternalkan ekspresi mereka sendiri tentang masalah target dalam bahasa alami. Hal ini dapat memperluas cakupan kueri kata kunci dan membuang sinonim untuk mendukung pengambilan bahasa alami. Kelas Jenis Tugas menjelaskan kategorisasi tugas berdasarkan fungsionalitas yang dapat diimplementasikan oleh tugas. Analisis MEW pada contoh di atas adalah sejenis tugas geografis. Kelas Tipe Tugas ditautkan ke kelas Tugas untuk referensi semantik guna menyatakan tipe tugas individu melalui properti hasTaskType yang telah ditentukan sebelumnya. Setiap individu dari kelas Task memiliki setidaknya satu solusi konseptual yang dilambangkan dalam ontologi Proses. Antarmuka kelas Task didefinisikan dalam Ontologi Antarmuka, yang akan dijelaskan secara rinci di bagian berikut.

Ontologi Proses

Ontologi Proses digunakan untuk mendefinisikan proses pemecahan masalah pada tingkat konseptual untuk jenis tugas tertentu, yang tidak terkait dengan layanan konkret apa pun. Kelas AtomicProcess dan CompositeProcess dibuat sebagai subkelas dari kelas Proses untuk mengklasifikasikan individu proses berdasarkan jumlah proses yang terlibat. Proses atom secara langsung mengacu pada kelas Geooperator dalam Ontologi Operasi GIS menggunakan properti RDF:Type; namun, proses komposit adalah himpunan tepi yang berisi beberapa tepi. Setiap tepi menunjukkan urutan dua node tugas yang dianotasi secara semantik ke kelas Tugas menggunakan properti fromTask dan toTask. Serangkaian sisi membentuk grafik berarah yang disebut grafik proses tugas yang menggambarkan cara kerja tugas. Dalam bab ini, kami hanya mempertimbangkan urutan linier antara dua tugas; logika aliran kontrol lainnya akan dimasukkan dalam pekerjaan masa depan.

Ontologi Tipe Data

Tipe Data Ontologi didefinisikan untuk mendeskripsikan tipe data yang dibagi menjadi dua kategori: SimpleDataType dan GeoDataType, seperti diilustrasikan pada Gambar 6. SimpleDataType menyertakan beberapa tipe data primitif dalam beberapa bahasa pemrograman atau bahasa deskripsi seperti xml:string dan xml:float in XML. GeoDataType adalah representasi abstrak data geografis, yang memiliki beberapa properti data yang dimiliki oleh semua jenis data geografis, termasuk atribut, format data, dan sistem referensi koordinat (CRS). Berdasarkan spesifikasi abstrak Organisasi Standar Internasional (ISO) untuk data vektor dan raster, GeoDataType dibedakan menjadi VectorDataType dan RasterDataType, yang masing-masing memiliki karakteristik unik. Dalam data vektor, setiap fitur geografis harus mengidentifikasi tipe geometris, seperti titik, polyline, dan poligon mengikuti Spesifikasi Fitur Sederhana OGC. Resolusi dan nomor pita harus diidentifikasi dalam data raster.



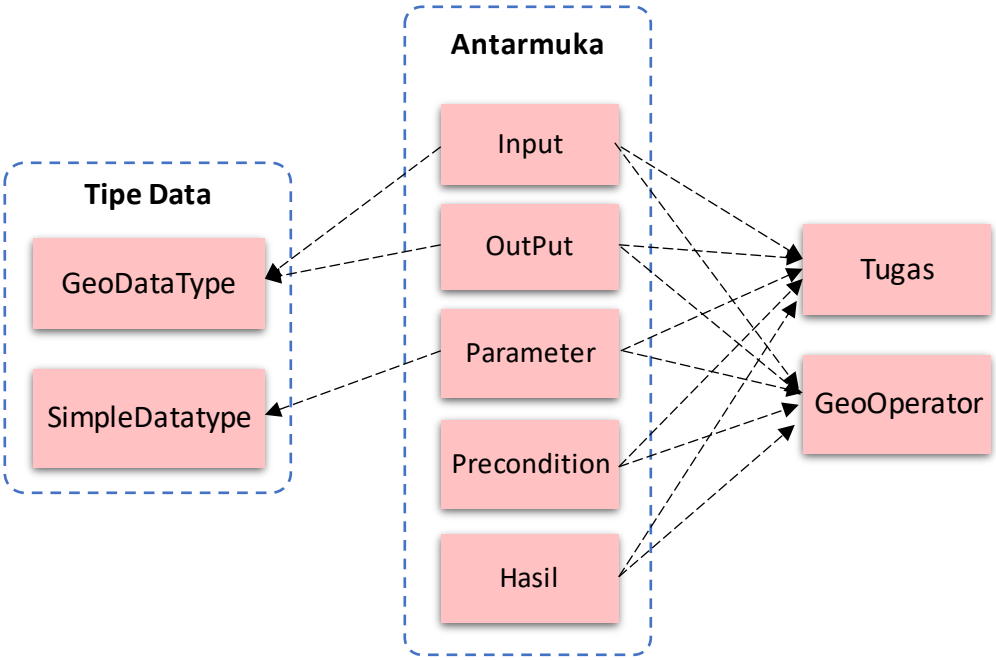
Gambar 9.6. Spesifikasi tipe data.

Ontologi Operasi GIS

Dalam Ontologi Operasi GIS, kelas Geooperator digunakan untuk mengkonsep fungsionalitas geoproses. Gagasan tentang Geooperator telah diperkenalkan di bagian sebelumnya. Operator geo digunakan sebagai blok bangunan untuk alur kerja konseptual pemecahan masalah geografis. Ontologi basis pengetahuan ini didasarkan pada karya Hofer yang menerjemahkan tesaurus SKOS (Sistem Organisasi Pengetahuan Sederhana) yang disediakan oleh Brauner ke dalam ontologi OWL dan memasukkan konsep tambahan yang dikenal sebagai konsep fungsional. Tesaurus SKOS berisi 40 geooperator. Ontologi ini dapat diperluas dengan kategori tambahan, jika perlu. Kategori-kategori dalam perspektif Pragmatis berasal dari tugas umum, dan merupakan kategori berorientasi tugas. Pengguna selanjutnya dapat mengintegrasikan kategori baru berdasarkan aplikasi praktis. Oleh karena itu, dalam bab ini, kategori tambahan bernama MEW diintegrasikan ke dalam perspektif Pragmatis geooperator, dan subkategori atau geooperator dapat dibuat untuk penjelasan lebih lanjut tentang operasi geoproses. Berdasarkan klasifikasi ini, layanan geoproses yang menjalankan fungsi geografis dianggap sebagai individu dari kelas Geooperator.

Ontologi Antarmuka

Seperti yang diperkenalkan pada bagian sebelumnya, tugas digunakan sebagai komponen yang dapat digunakan kembali untuk mencapai komposisi proses pemecahan masalah. Komposisi memerlukan evaluasi korespondensi antarmuka. Basis pengetahuan perlu mencakup informasi antarmuka yang memadai untuk memenuhi kebutuhan komposisi. Sebuah antarmuka memerlukan deskripsi operan yang berisi input dan output, batasan yang berisi prasyarat dan hasil, serta parameter non-spasial. Akibatnya, seperti yang diilustrasikan pada Gambar 9.7, kelas Antarmuka terdiri dari subkelas Input, Output, Parameter, Prakondisi, dan Hasil. GeoDataType dalam Ontologi Tipe Data digunakan untuk menentukan operan antarmuka, sedangkan parameter non-spasial dapat merujuk pada SimpleDataType yang mencakup tipe data konvensional. Kelas Precondition berfokus pada sifat tematik dan geometris dari input untuk memastikan fungsi operasi yang benar. Kelas Postcondition mendefinisikan hasil keluaran yang diharapkan.



Gambar 9.7. Antarmuka untuk anotasi Tugas dan Geooperator, serta Tipe Data untuk menentukan Antarmuka.

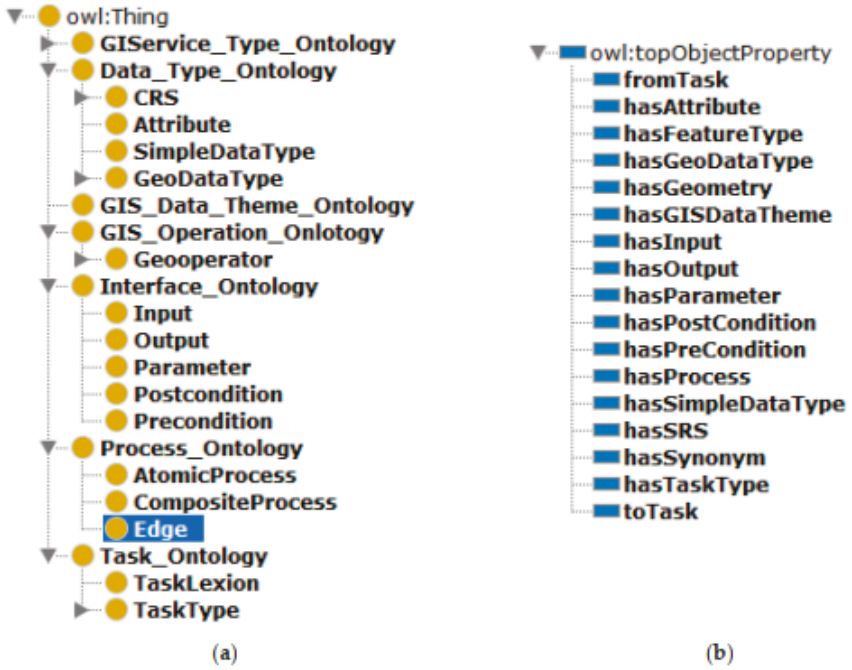
Demikian pula, kami memperluas properti antarmuka geooperator menggunakan Ontologi Antarmuka yang saat ini tidak melibatkan spesifikasi antarmuka terkait.

9.6 IMPLEMENTASI SISTEM GEOGRAFIS

Pada Bagian sebelumnya kami memperkenalkan skenario aplikasi yang merupakan proses penyelesaian masalah geografis dalam konteks MEW. Kami mengambil contoh ini untuk menunjukkan manfaat dari basis pengetahuan berbasis ontologi untuk tugas-tugas selama proses pemecahan masalah geografis. Implementasinya mencakup tiga bagian: pembuatan ontologi, representasi pengetahuan, dan contoh tugas.

Penciptaan Ontologi

Berdasarkan arsitektur yang diusulkan dari basis pengetahuan berorientasi tugas yang dijelaskan dalam Bagian sebelumnya, kami membangun ontologi abstrak yang berbeda untuk mewakili hierarki dan hubungan konsep menggunakan Protégé 5.2.0 yang merupakan platform pengembangan ontologi OWL yang memungkinkan pembuatan dan kueri ontologi. Secara umum, ontologi terdiri dari komponen-komponen berikut: konsep dan properti setiap konsep, hubungan atau batasan antar konsep, dan contoh konsep. Gambar 9.8a menyajikan semua konsep atau kelas yang didefinisikan dalam basis pengetahuan ontologis. Seluruh properti objek yang mewakili hubungan antar kelas ditunjukkan pada Gambar 8b; mereka termasuk hasTaskType, hasSynonym, hasProcess, dll. Ontologi abstrak dapat dipakai untuk tugas-tugas tertentu. Dalam tulisan ini, contoh-contoh tugas peringatan dini meteorologi diimplementasikan, yang akan dirinci pada bagian berikutnya.



Gambar 9.8. Cuplikan ontologi dimana (a) menggambarkan kelas-kelas ontologi, dan (b) menggambarkan properti objek antar kelas.

Representasi Pengetahuan Ontologi

Setelah komponen ontologi dikembangkan, ontologi tersebut dapat direpresentasikan dengan bahasa deskripsi ontologi, seperti Resource Description Framework (RDF) dan Web Ontology Language (OWL). RDF dibangun berdasarkan XML, yang menggunakan tiga kali lipat objek, properti, dan nilai untuk mendeskripsikan sumber daya. OWL adalah bahasa markup semantik standar yang direkomendasikan W3C yang dikembangkan oleh komunitas Web Semantik, yang merupakan perpanjangan dari RDF. Dalam buku ini, kami menggunakan OWL sebagai bahasa standar dan dapat dibaca mesin untuk mewakili pengetahuan ontologi, yang disajikan sebagai file OWL.

Sementara itu, kami menggunakan batasan properti termasuk batasan hasValue dan quantifier untuk membatasi asosiasi antar kelas yang berbeda. Pembatasan hasValue menentukan bahwa individu dalam suatu kelas mempunyai nilai tertentu. Namun demikian, pembatasan kuantifikasi membatasi individu dalam suatu kelas menggunakan pembatasan eksistensial (∃, burung hantu:someValuesFrom) atau pembatasan universal (∀, burung hantu:allValuesFrom). Yang pertama menyatakan bahwa nilai-nilai untuk properti yang dibatasi memiliki setidaknya satu contoh kelas, yang ditentukan oleh pembatasan eksistensial; namun, yang terakhir menyatakan bahwa semua nilai untuk hubungan terbatas harus berupa tipe instance. Misalnya, tugas analisis MEW hanya memerlukan data curah hujan yang efektif, data prakiraan curah hujan, dan data derajat potensial yang dapat dibatasi dengan pernyataan formal berikut: ∀ hasInput (Data_Hujan_Efektif ∪ Data_Hujan_Perkiraan ∪ Data_Derajat_Potensial). Pernyataan ini mendefinisikan batasan universal pada properti “hasInput” antara kelas Task dan kelas Input (Gambar 9.5). Notasi OWL yang menggunakan batasan “owl:allValuesFrom” ditunjukkan pada Gambar 9.9.

```
<owl:Class rdf:ID = "Meterological Early-warning Task " >
        <rdfs:subClassof>
                <owl:Restriction>
                        <owl:onProperty rdf:resource = "hasInput " />
                        <owl:allValuesFrom rdf:resource = "#Effective_Rainfall_Data " />
                        <owl:allValuesFrom rdf:resource = "#Forecast_Rainfall_Data " />
                        <owl:allValuesFrom rdf:resource = "#Potential_Degree_Data " />
                </owl:Restriction>
        </rdfs:subClassof>
</owl:Class>
```

Gambar 9.9. Cuplikan notasi OWL menggunakan batasan universal.

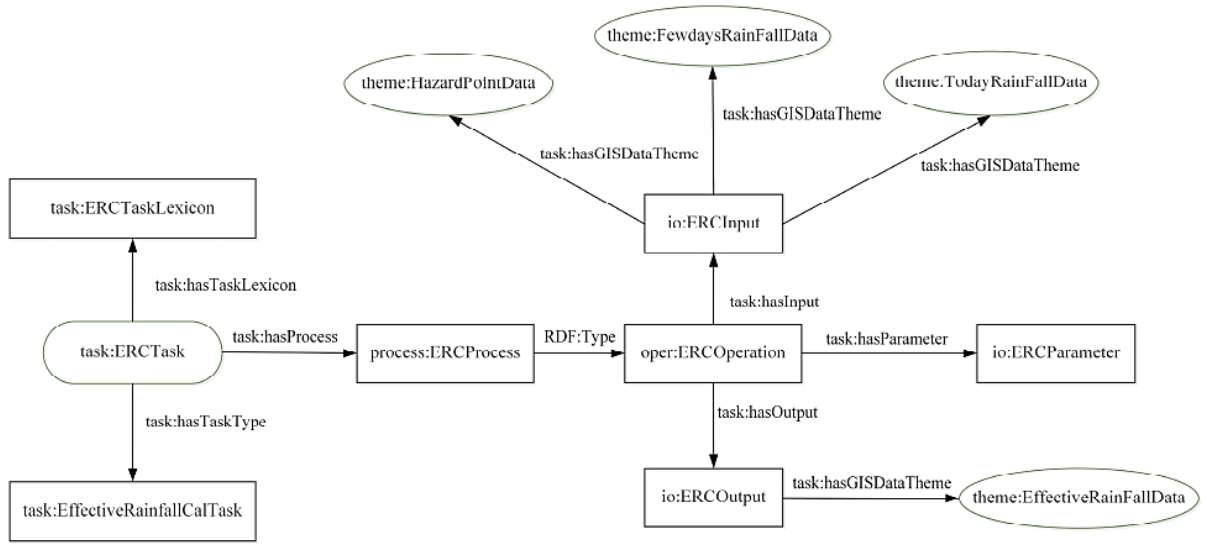
Contoh tugas tertentu dapat direpresentasikan menggunakan kelas dan properti yang ditentukan dalam ontologi. Dengan menggunakan peringatan dini meteorologi sebagai contoh, tugas-tugas yang terlibat dalam contoh MEW tercantum dalam Tabel 9.1, yang di dalamnya terdapat dua tugas gabungan (misalnya, EWATask) dan enam tugas atom (misalnya, ERCTask, dan FQTask). Kami menggunakan ERCTask sebagai contoh contoh tugas atom, yang digunakan untuk menghitung curah hujan efektif. Gambar 9.10 menunjukkan individu dan properti yang terlibat dalam contoh ERCTask. Proses tugas atom merupakan bagian dari AtomicProcess, sedangkan tugas gabungan tidak. Kami mencantumkan deklarasi namespace ontologi dan sintaks definisi kelas, subkelas, dan properti menggunakan OWL, seperti yang ditunjukkan di bawah Gambar 9.10.

Tabel 9.1. Tugas-tugas yang terlibat dalam contoh MEW

Jenis Tugas	Singkatan	SubTugas	Keterangan
Potensi gelar callask	Tugas PDC	FQ tugas FWC tugas PDI tugas	Hitung indeks derajat potensial dari beberapa data faktor pegaruh
Tugas panggilan curah hujan detektif	Tugas ERC		Hitung curah hujan efektif
Tugas analisis peringatan dini	Permintaan EWA	QA Tugas HRI tugas EWL tugas	Menghasilkan peta prakiraan berdasarakan model peringatan dini
Tugas kuantifikasi faktor	Tugas FQ		Hitung data faktor menurut model faktor kepastian
Panggilan Tugas faktor bobot	Tugas FWCT		Hitung bobot faktor
Panggilan Tugas potensi gelar index	Tugas PDI		Hitungan indeks derajat potensial
Tugas analaisis everlay	Tugas OAT		Overlay data masukan
Panggilan tugas Indeks resiko bahaya	Tugas HR		Hitung indeks risiko bahaya

Tingkat peringatan dini	Tugas EWL	Beginilah Tingkat peringatan dini berdasarakan indeks resiko

Berbeda dari tugas atomik, proses tugas gabungan terdiri dari beberapa individu tepi, yang masing-masing menggambarkan aliran data antara dua contoh tugas. Sekumpulan sisi menyusun grafik proses yang menunjukkan cara kerja tugas. Misalnya, Gambar 9.11 menunjukkan contoh tugas dari tugas gabungan yang disebut EWATask. Proses individu "proses:EWAProcess" berisi dua individu tepi: "proses:EWAEdge1" dan "proses:EWAEdge2". Yang pertama menghubungkan dua contoh tugas: "task:QATask" dan "task:HRITask", dan yang terakhir menghubungkan "task:HRITask" dan "task:EWLTask". Individu edge ini ditautkan ke individu proses dengan properti itemEdge.




```
###Define namespace declaration of ontology
<rdf:RDF
    xmlns = "http://www.geojie.com:8089/ontologies/tasks/task.owl#"
    xmlns:owl = "http://www.w3.org/2002/07/owl#"
    xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs = "http://www.w3.org/2000/01/rdf-schema#"
    xmlns:xsd = "http://www.w3.org/2001/XMLSchema#"
    xmlns:task = "http://www.geojie.com:8089/ontologies/tasks/task.owl#"
    xmlns:process = "http://www.geojie.com:8089/ontologies/tasks/process.owl#"
    xmlns:io = "http://www.geojie.com:8089/ontologies/tasks/interface.owl#"
    xmlns:oper = "http://www.geojie.com:8089/ontologies/tasks/geooperator.owl#"
    xmlns:theme ="http://www.geojie.com:8089/ontologies/tasksdatatheme.owl#">
```

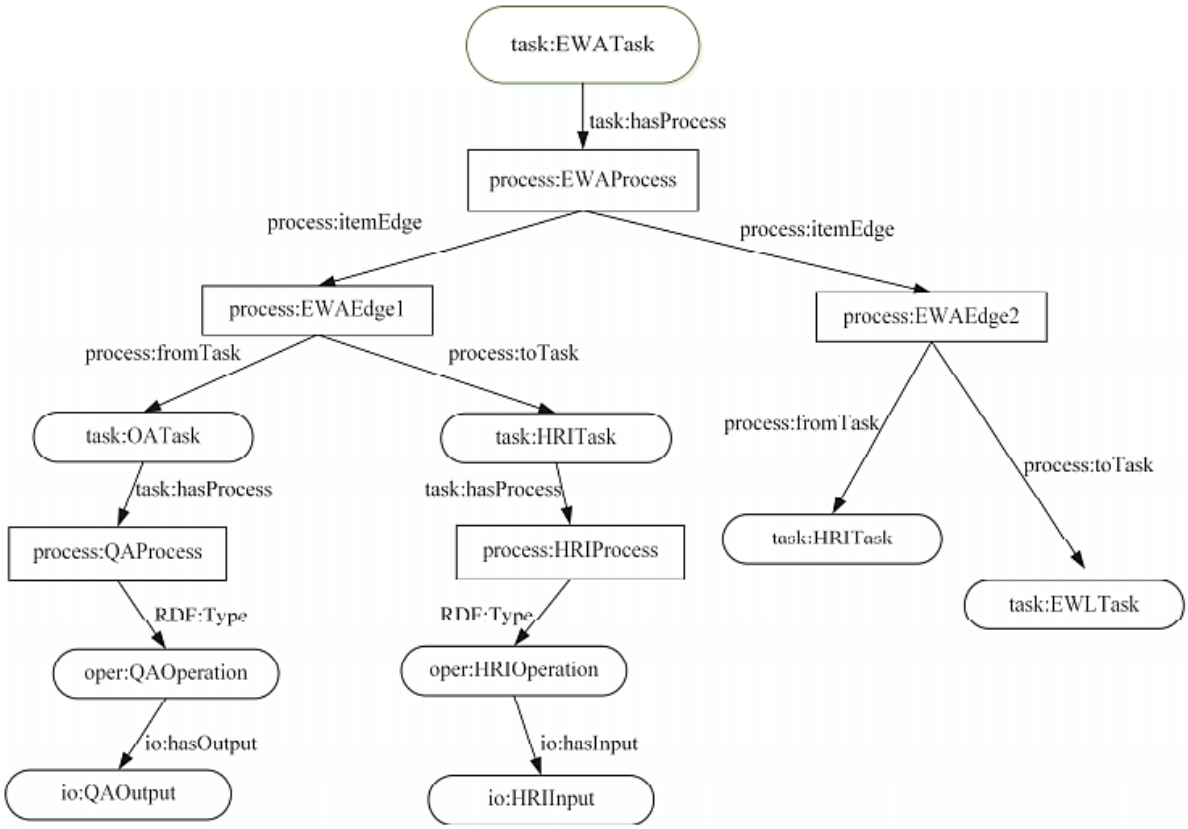
```
###Define named class
<owl:Class rdf:ID="Task"/>

###Define subclass
<owl:Class rdf:ID="TaskType">
    <rdfs:subClassOf rdf:resource="#Task" />
</owl:Class>

###Define individual of class
<Task rdf:ID="ERCTask"/>

###Define properties between classes
<owl:ObjectProperty rdf:ID="hasTaskType">
  <rdfs:domain rdf:resource="#Task"/>
  <rdfs:range rdf:resource="#TaskType"/>
</owl:ObjectProperty>
```

Gambar 9.10. Contoh tugas dari tugas atom (EffectiveRainfallCalTask).



Gambar 9.11. Contoh tugas dari tugas gabungan (EarlyWarningAnalysisTask).

Prototipe

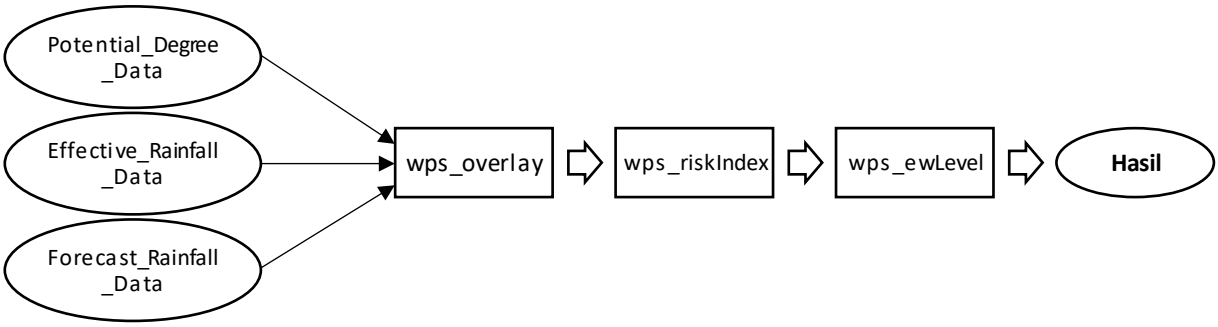
Sebuah sistem prototipe berdasarkan representasi ontologi yang direalisasikan dan contoh tugas yang diformalkan dikembangkan untuk memfasilitasi pengguna dalam memecahkan masalah geografis yang kompleks. Prototipe yang diimplementasikan memanfaatkan sejumlah teknik web, seperti Ajax, XML, JSON, EasyUI, GoJS, OpenLayers, Apache Axis2, dan sebagainya. Ajax digunakan untuk pertukaran data asinkron antara sisi klien dan server. XML dan JSON adalah format pertukaran data. EasyUI dan GoJS adalah kerangka UI klien, dan GoJS digunakan untuk menggambar diagram alur. OpenLayers adalah paket perpustakaan kelas JavaScript untuk pengembangan klien WebGIS, yang digunakan untuk mencapai akses data peta. Apache Axis2 digunakan untuk menyediakan antarmuka Layanan Web. Paket Java API Apache Jena, kerangka Web Semantik untuk Java, digunakan untuk mengurai file ontologi, mengakses definisi ontologi, dan menyimpulkan pengetahuan. Server Apache Tomcat digunakan sebagai wadah web. Sistem prototype dapat diakses menggunakan Microsoft IE atau browser Google pada sistem operasi Windows.

Analisis MEW di Henan, China dijadikan sebagai contoh pemanfaatan basis pengetahuan untuk mendukung penyelesaian masalah geografis. Pertama, kami mendefinisikan semantik formal dalam basis pengetahuan berbasis ontologi dengan membuat contoh tugas menggunakan editor ontologi. Contoh tugas diberi nama EWATask (Gambar 9.11) yang dapat didekomposisi menjadi tiga subtugas (OATask, HRITask, dan EWLTask), File ontologi dihasilkan menggunakan bahasa format OWL yang disebutkan di bagian sebelumnya.

Kedua, layanan web, termasuk tiga layanan akses data (Potential_Degree_Data, Effective_Rainfall_Data, dan Forecast_Rainfall_Data) dan tiga layanan geoproses (wps_overlay, wps _riskIndex, dan wps_ewLevel), diterbitkan dengan dukungan MapGIS IGServer. Rincian layanan akses data ditunjukkan pada Tabel 9.2, layanan geoproses mengikuti spesifikasi WPS, dan model alur kerja untuk EWATask ditunjukkan pada Gambar 9.12.

Tabel 9.2. Layanan akses data yang terlibat dalam EWATask.

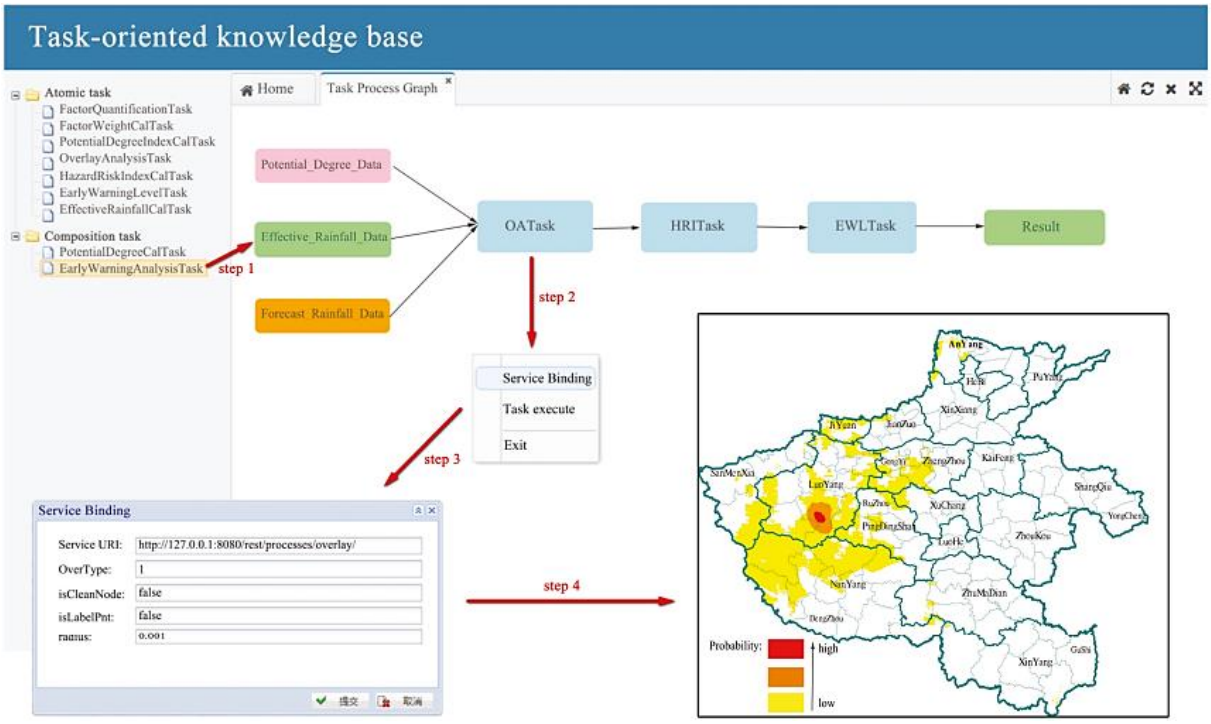
Nama Data	Jenis Layanan	SRS	Geometri
Potential_Degree_Data	WFS	Xi'an 80	Poligon
Effective_Rainfall_Data	WFS	Xi'an 80	Poligon
Forecast_Rainfall_Data	WFS	Xi'an 80	Poligon



Gambar 9.12. Model alur kerja EWATask. Layanan akses data direpresentasikan dalam bentuk elips, dan layanan geoproses direpresentasikan dalam bentuk persegi panjang.

Terakhir, sistem prototipe menyediakan antarmuka pengguna grafis (GUI) yang intuitif dan mudah digunakan. Pengguna akhir dapat mengakses GUI sistem prototipe menggunakan browser web. Seperti yang ditunjukkan pada Gambar 9.13, di panel kiri, terdapat struktur pohon yang memperlihatkan daftar tugas yang diurai dari basis pengetahuan. Pengguna memilih dan mengklik simpul tugas; kemudian model proses tugas akan ditampilkan dalam bentuk diagram alur di panel kanan (langkah 1). Selanjutnya, klik kanan pada node proses dan pilih menu "service binding" (langkah 2). Jendela pengikatan layanan muncul dan memungkinkan pengguna akhir untuk mengikat layanan yang sesuai dan memasukkan parameter terkait secara manual (langkah 3). Ulangi langkah ini untuk setiap node proses. Terakhir, jalankan tugas dan dapatkan peta hasil (langkah 4). Misalnya, Wang menjadikan provinsi Henan di Tiongkok sebagai wilayah perkiraan untuk analisis risiko dan memperkirakan kemungkinan terjadinya bahaya geologi dalam 24 jam ke depan. Dia mengklik node EarlyWarnAnalysisTask di sistem prototipe, yang grafik prosesnya ditunjukkan di panel kanan (Gambar 9.13). Mengikuti alur kerja ini, dia mengikat layanan geoproses yang sesuai (wps_overlay, wps_riskIndex, dan wps_ewLevel) yang dipanggil dengan urutan linier (wps_overlay → wps_riskIndex → wps_ewLevel). Berdasarkan hasil prakiraan, diperoleh peta

hasil peringatan dini seperti yang ditunjukkan di kanan bawah Gambar 9.13, yang menggunakan warna berbeda untuk mewakili tingkat peringatan dini yang berbeda.



Gambar 9.13. Antarmuka pengguna grafis dari sistem prototipe.

9.7 RINGKASAN

Pada bab ini mengusulkan model tugas dan mengabstraksi masalah geografis sebagai tugas yang dapat digunakan sebagai komponen yang dapat digunakan kembali untuk pemecahan masalah. Basis pengetahuan berorientasi tugas dibangun untuk menangkap pengetahuan pemecahan masalah geografis yang dapat dibagikan dan digunakan kembali. Dalam basis pengetahuan, kami menggabungkan beberapa ontologi (misalnya, Ontologi Tugas, Ontologi Proses, dan Ontologi Operasi GIS) untuk memberikan bantuan pada semua aspek pemecahan masalah. Basis pengetahuan ini tidak terkait erat dengan bahasa alur kerja tertentu. Pengetahuan yang diperlukan tentang pemecahan masalah disimpan dalam basis pengetahuan yang menggunakan ontologi dan pendekatan berorientasi tugas untuk mencapai formalisasi dan penggunaan kembali tugas.

Basis pengetahuan ini dirancang bagi para pakar domain untuk membuat dan berbagi pengetahuan pemecahan masalah geografis profesional mereka. Bagi pengguna akhir, antarmuka yang ramah pengguna diperlukan untuk menyampaikan masalah geografis dan menanyakan solusi masalah. Suatu pendekatan yang memiliki kemampuan mengurai masukan bahasa alami akan dikembangkan dalam penelitian mendatang. Pendekatan ini akan memungkinkan pengguna memasukkan teks bebas untuk mengajukan persyaratan masalah.

BAB 10
GRAFIK PENGETAHUAN GEOGRAFIS (GEOKG)

Representasi pengetahuan yang diformalkan adalah dasar dari komputasi, penambangan, dan visualisasi Big Data. Representasi pengetahuan saat ini menganggap informasi sebagai item yang dihubungkan dengan objek atau konsep yang relevan melalui struktur pohon atau grafik. Namun pengetahuan geografis berbeda dengan pengetahuan umum yang lebih terfokus pada pengetahuan temporal, spasial, dan perubahan. Dengan demikian, item pengetahuan terpisah sulit untuk mewakili keadaan geografis, evolusi, dan mekanisme, misalnya, proses badai "{9:30-60 mm-presipitasi}-{12:00-80 mm-presipitasi}-... ".

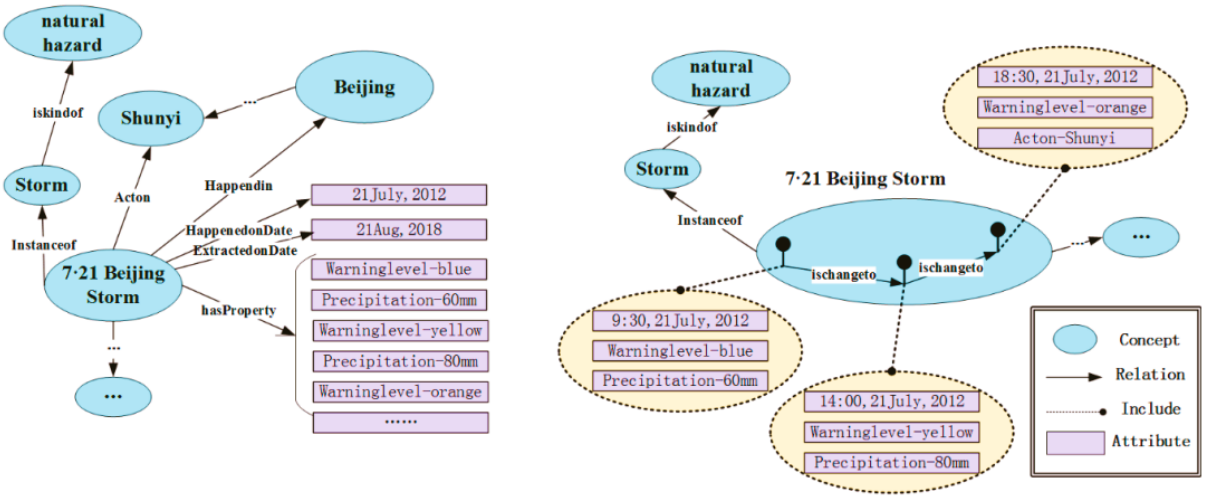
Masalah mendasarnya adalah konstruktor landasan logika (bahasa deskripsi ALC) dari representasi pengetahuan geografis saat ini, yang tidak dapat memberikan deskripsi tersebut. Untuk mengatasi masalah ini, penelitian ini merancang representasi pengetahuan geografis formal yang disebut GeoKG dan melengkapi konstruktor bahasa deskripsi ALC. Kemudian, kasus evolusi pembagian administratif Nanjing diwakilkan kepada GeoKG. Untuk mengevaluasi kemampuan model formal kami, dua grafik pengetahuan dibuat dengan menggunakan GeoKG dan YAGO dengan menggunakan kasus pembagian administratif. Kemudian, serangkaian pertanyaan geografis didefinisikan dan diterjemahkan ke dalam pertanyaan. Hasil query menunjukkan bahwa hasil GeoKG lebih akurat dan lengkap dibandingkan YAGO dengan informasi keadaan yang disempurnakan. Selain itu, evaluasi pengguna memverifikasi peningkatan ini, yang menunjukkan bahwa ini adalah model yang menjanjikan dan kuat untuk representasi pengetahuan geografis.

10.1 PENDAHULUAN

Pengetahuan geografis terdiri dari produk pemikiran dan penalaran geografis tentang fenomena alam dan manusia di dunia, yang memainkan peran penting dalam studi dan penerapan geografi. Hampir setiap ahli geografi mencoba menjawab pertanyaan "bagaimana memahami, memahami, dan mengatur pengetahuan geografis secara ilmiah." Secara umum, representasi pengetahuan geografis adalah jenis ekspresi manusia terhadap dunia nyata yang sangat penting untuk penyimpanan dan komputasi. Khususnya di era Big Data, pengetahuan geografis yang terstruktur dengan baik merupakan manfaat bagi semua jenis aplikasi geografis, karena formalisasi adalah dasar dari komputasi, penambangan, dan visualisasi big data geografis.

Saat ini, representasi pengetahuan yang paling populer adalah grafik pengetahuan. Ini mengatur pengetahuan dengan sekumpulan konsep, relasi, dan fakta, yang diasosiasikan oleh dua tipe {entitas, relasi, entitas} dan {entitas, atribut, nilai atribut}. Hanya ada tiga elemen dasar dalam grafik pengetahuan: entitas, relasi, dan atribut. Ketiga elemen ini secara eksplisit dapat mewakili informasi umum, seperti "kapan badai Beijing terjadi pada 21 Juli—09:30, 21 Juli". Namun, pengetahuan geografis lebih rumit dibandingkan pengetahuan umum. Lebih banyak proses dan evolusi yang perlu dijawab, misalnya, "apa yang menyebabkan badai 7·21

Beijing", "bagaimana perkembangannya", dan "apa dampak dari badai 7·21 Beijing". Entitas, relasi, dan atribut tidak dapat dengan mudah dan langsung menjawab pertanyaan-pertanyaan mekanika ini. Misalnya, representasi grafik pengetahuan geografis dari badai Beijing 7·21 ditunjukkan pada Gambar 10.1.



Gambar 10.1. Representasi pengetahuan geografis yang berbeda dari Badai Beijing 7·21. (a) Struktur data grafik pengetahuan dan (b) struktur data pengetahuan prosedural.

Gambar 10.1a mengatur pengetahuan geografis badai Beijing 7·21 menggunakan struktur data grafik pengetahuan saat ini. Model representasi pengetahuan ini dapat secara eksplisit merepresentasikan setiap fakta dan relasinya. Namun, hal ini tidak dapat mewakili evolusi atau mekanisme, yang merupakan topik utama dalam geografi. Selain itu, jenis representasi pengetahuan ini sangat berbeda dari struktur data pengetahuan prosedural yang ditunjukkan pada Gambar 9.1b. Secara umum manusia mempersepsikan objek, peristiwa, dan aktivitas melalui pengolahan pengetahuan deklaratif, pengetahuan prosedural, dan pengetahuan struktural. Dan pengetahuan prosedural memberikan kerangka kerja pada pengetahuan deklaratif, yang bermanfaat untuk memahami mekanisme yang mendasarinya. Badai 7·21 Beijing meliputi tiga tahap utama: 09:30, 14:00, dan 18:30. Setiap tahap memiliki daftar atribut. Struktur data pengetahuan prosedural ini membantu orang mengetahui evolusi atau mekanisme secara lebih eksplisit. Misalnya, masyarakat tidak dapat secara langsung memahami bahwa semua atribut (tingkat peringatan biru, tingkat peringatan kuning, dll.) terkait dengan "badai Beijing 7·21", padahal masyarakat dapat mengetahui bahwa badai tersebut memiliki tingkat peringatan yang berbeda-beda pada tahapan yang berbeda.

Tujuan dari pembahasan bab ini adalah untuk meningkatkan grafik fakta pengetahuan diskrit deklaratif menjadi pengetahuan agregat prosedural. Untuk mengatasi masalah ini, bab ini menyajikan model formal untuk representasi pengetahuan geografis dari perspektif geografi, yang disebut GeoKG, dan melengkapi konstruktor bahasa deskriptif ALC.

Sisa dari bab ini disusun mengulas karya-karya terkait ontologi geografis dan grafik pengetahuan geografis. Serta menjelaskan metodologi dengan menyatakan ide-ide dasar dari enam pertanyaan inti geografis dan mengusulkan model representasi pengetahuan geografis

yang diformalkan yang disebut GeoKG. Studi kasus evolusi pembagian administratif Nanjing dengan model GeoKG yang diformalkan. Kemudian menyusun kasus pembagian administrasi dengan menggunakan model GeoKG dan model YAGO, menetapkan serangkaian pertanyaan, dan menganalisis hasilnya. Terakhir menyajikan kesimpulan.

10.2 ONTOLOGI GEOGRAFIS

Ontologi geografis berasal dari ontologi yang mewakili teori-teori filosofis paling dasar yang mewakili sifat dan karakteristik dunia nyata. Pada tahun 1960an, "ontologi" diperkenalkan dalam ilmu informasi untuk kategorisasi, representasi, berbagi pengetahuan, dan penggunaan kembali. Ontologi geografis adalah konsep domain, yang secara eksplisit dan formal mendefinisikan konsep geografis dan hubungannya dalam geografi melalui hubungan hierarki. Hubungan hierarki antar konsep ini penting bagi representasi pengetahuan geografis, integrasi informasi, interoperasi pengetahuan, dan pengambilan informasi. Dengan demikian, ontologi geografis merupakan metode representasi pengetahuan geografis penting yang diterapkan secara luas dalam berbagai aplikasi informasi geografis. Namun, simulasi komputer tidak hanya memerlukan logika konsep hierarki standar tetapi juga sejumlah besar informasi contoh dalam representasi pengetahuan geografis. Ada dua jenis ontologi geografis yang dibatasi oleh representasi pengetahuan geografis.

Pertama, ontologi geografis berfokus pada struktur sistem konseptual, yang dibangun secara ketat informasi hiponim. Hubungan ini cocok untuk kategorisasi, disambiguasi, identifikasi dan inferensi tetapi tidak untuk menggambarkan keadaan dan fenomena perubahan objek geografis. Deskripsi perubahan keadaan dan fenomena ini memerlukan banyak informasi, yang tidak terdapat dalam ontologi geografis. Meskipun hiponimi didefinisikan secara ketat oleh struktur pohon hierarki dalam ontologi geografis, struktur ini tidak dapat secara langsung mewakili hubungan antara berbagai konsep yang penting untuk mewakili evolusi dan mekanisme dalam geografi. Selain itu, hubungan antar simpul dalam sebuah pohon tidak bersifat dua arah sehingga membatasi representasi interaksi antar objek geografis. Penyebab permasalahan tersebut terkait dengan struktur pohon yang membatasi representasi pengetahuan geografis.

Kedua, landasan logika ontologi geografis adalah logika deskripsi (DL) bahasa konsep atributif dengan pelengkap (ALC). DL adalah bahasa representasi pengetahuan formal berbasis objek. Ini berisi empat komponen: satu set konstruksi mewakili konsep dan peran (misalnya, sungai adalah sebuah konsep; disjungsi adalah sebuah peran), pernyataan tentang konsep terminologi (kotak terminologi, Tbox, misalnya, setiap sungai memiliki panjangnya sendiri), pernyataan tentang item individual (kotak pernyataan, Abox, misalnya, panjang Sungai Changjiang adalah 6300 km) dan mekanisme penalaran Tbox dan Abox. DL dapat membangun konsep dan peran yang rumit dengan konsep dan peran sederhana oleh konstruktor. Menurut konstruktor yang berbeda, DL dapat diklasifikasikan menjadi ALC, ALCN, S, SH, SHIQ, dll. ALC adalah DL dasar yang berisi perpotongan ($\sqcap$), gabungan ($\sqcup$), komplemen ($\neg$), batasan universal ($\forall$), dan batasan eksistensial ($\exists$). ALCN terdiri dari operator ALC dasar dan batasan angka (Q; $\geq$ n dan $\leq$ n); ALC+R , kependekan dari S, terdiri dari deskripsi dasar dan operator hubungan

yang ditingkatkan ® ; transisi peran atau konsep); bahasa SH, dengan inklusi konsep dan inklusi peran; dan SHIQ mencakup peran terbalik (I) dan transisi peran (R+). Saat ini, logika deskripsi SHIQ telah disertifikasi untuk mewakili perubahan dalam bidang teori logika. Perhatikan bahwa "perubahan" adalah elemen yang sangat penting untuk representasi pengetahuan geografis dan berarti bahwa konstruktor ALC tidak dapat mewakili semua hubungan logis dari pengetahuan geografis, terutama dalam ekspresi kuantitas dan perubahan keadaan. Misalnya, konstruktor pembatasan jumlah diharuskan untuk mewakili pengetahuan geografis "Sungai Yangzi memiliki setidaknya tiga cabang" (∃ memiliki cabang; Sungai Yangzi ≥ 3), dan konstruktor transisi diharuskan untuk mewakili pengetahuan geografis "Beiping diubah namanya menjadi Beijing pada 27 September 1949" (Beiping ≡ Trans(Beijing)). Sementara itu, banyak penelitian secara teoritis mendemonstrasikan dan membuktikan decidability, soundness dan kelengkapan operator rangkaian DL (dari ALC, S, SI, SHI hingga SHIQ, dll.) pada algoritma Tableau, dan kompleksitasnya. SI (transisi peran atau konsep) adalah PSPACE lengkap dan SHI dan SHIQ berikut adalah EXPTIME lengkap.

10.3 GRAFIK PENGETAHUAN GEOGRAFIS

Grafik pengetahuan merupakan model representasi pengetahuan yang berbentuk grafik dengan logika yang ketat, konsep yang berbeda, relasi yang beragam, dan instance yang masif. Ini pertama kali disajikan oleh Google pada tahun 2012, berisi lebih dari 5,7 miliar entitas dan 0,18 miliar fakta. Dengan kekayaan informasi ini, dunia nyata dapat digambarkan secara eksplisit. Penyimpanan berbasis grafik memiliki sifat koneksi, arah, dan multi-simpul yang cocok untuk merepresentasikan interaksi antar konsep. Dengan demikian, grafik pengetahuan merupakan model yang menjanjikan untuk mewakili pengetahuan dan telah dibangun secara luas, misalnya YAGO, Freebase, Probase, dan DBpedia. Grafik pengetahuan geografis merupakan grafik pengetahuan domain yang sedang dalam tahap eksplorasi.

Saat ini, sebagian besar grafik pengetahuan geografis disusun sebagai grafik pengetahuan universal, misalnya CSGKB, NCGKB, dan CrowdGeoKG. Basis pengetahuan geografis akal sehat (CSGKB) menggunakan struktur data yang menghubungkan konsep fitur geografis, lokasi geografis, hubungan spasial, dan administrator pengambilan informasi geografis (GIR) alih-alih pengukuhan tradisional. Selain itu, basis pengetahuan geografis Tiongkok yang naif (NCGKB) membangun basis pengetahuan geografis berorientasi GIR berdasarkan Wikipedia bahasa Mandarin berdasarkan hubungan konsep tertentu dan contohnya. CrowdGeoKG menggunakan grafik pengetahuan geografis crowdsourced yang mengekstrak berbagai jenis entitas geografis dari OpenStreetMap dan memperkayanya dengan informasi geografi manusia dari Wikidata. Semua konsep grafik pengetahuan geografis ini dikembangkan berdasarkan ontologi geografis yang mengikuti bahasa deskriptif ALC, sehingga menimbulkan permasalahan yang sama dengan ontologi geografis.

Lebih penting lagi, tiga basis saat ini mengatur pengetahuan geografis sebagai sekumpulan konsep, relasi, dan fakta, yang diasosiasikan oleh dua jenis tipe {entitas, relasi, entitas} dan {entitas, atribut, nilai atribut}. Sebenarnya hanya ada tiga elemen dasar dalam grafik pengetahuan: entitas, relasi, dan atribut. Ketiga elemen ini secara eksplisit dapat

mewakili informasi umum seperti "kapan 7·21 badai Beijing — 09:30, 21 Juli". Namun, pengetahuan geografis lebih rumit dibandingkan pengetahuan umum. Lebih banyak proses dan evolusi yang perlu dijawab, misalnya, "apa yang menyebabkan badai 7·21 Beijing", "bagaimana perkembangannya", dan "apa dampak dari badai 7·21 Beijing". Entitas, relasi, dan atribut tidak dapat dengan mudah dan langsung menjawab pertanyaan-pertanyaan mekanika ini.

Para ahli mengindikasikan bahwa dibutuhkan lebih banyak elemen. PLUTO melengkapi elemen waktu dengan "sebelum" dan "sesudah" dalam model grafik pengetahuan untuk menggambarkan lintasan perubahan objek geografis. Grafik pengetahuan geologi telah diterapkan dengan elemen evolusi untuk menyatakan perubahan antara objek geologi yang berbeda. YAGO juga mengeksplorasi penjangkaran dimensi spasial dan temporal ke basis pengetahuan, yang disebut YAGO2. YAGO2 membiarkan titik waktu dan interval waktu dengan format standar untuk menggambarkan informasi temporal dan mengatur koordinat geografis yang terkait dengan entitas untuk melengkapi informasi spasialnya. Faktanya, pengetahuan spasial dan temporal yang disimpan dalam sistem YAGO hanya dianggap sebagai atribut umum dengan menambahkan predikat seperti "wasBornOnDate", "occurs Since", "hasGeoCoordinates", dll., sedangkan informasi diskrit deklaratif tidak dapat langsung menjawab pertanyaan selanjutnya, evolusi dan mekanisme. Selain itu, sepuluh konsep inti ilmu informasi geografis diusulkan untuk penelitian transdisipliner: lokasi, lingkungan, bidang, objek, jaringan, peristiwa, granularitas, akurasi, makna, dan nilai. Konsep-konsep ini dapat mencakup setiap sudut geosains, namun sangat sulit untuk dihubungkan dengan satu model yang dikonseptualisasikan. Baru-baru ini, enam faktor (semantik geografis, lokasi, bentuk, proses evolusi, hubungan antar elemen, dan atribut) diusulkan untuk menggambarkan informasi dari elemen geografis, objek, atau fenomena. Meskipun faktor-faktor ini dirancang untuk representasi informasi objek geografis, faktor-faktor ini juga dapat memberikan panduan untuk representasi pengetahuan geografis. Dan semua penelitian di atas menunjukkan bahwa pengetahuan geografis dapat direpresentasikan secara lebih efektif dengan melengkapi elemen-elemennya, sekaligus memunculkan pertanyaan mendasar: "bagaimana mengorganisasikan pengetahuan geografis secara ilmiah dan kognitif?" Oleh karena itu, model grafik pengetahuan geografis yang dikonseptualisasikan dari perspektif geografi memerlukan studi lebih lanjut.

10.4 IDE PENCARIAN

Untuk mengatasi permasalahan di atas, pertanyaan inti model GeoKG adalah menentukan jenis pengetahuan geografis yang perlu disimpan. Geografi (dari bahasa Yunani γεωγραφία, geographia, secara harfiah berarti "deskripsi bumi") adalah bidang ilmu yang ditujukan untuk mempelajari daratan, fitur, penghuni, dan fenomena Bumi. Sebagai salah satu bentuk pemahaman manusia terhadap lingkungan geografis, pengetahuan geografis harus menjawab pertanyaan tentang geografi. Pertanyaan-pertanyaan tentang geografi telah dipisahkan menjadi enam pertanyaan inti oleh International Geographical Union (IGU), yang merupakan bagian dari piagam internasional tentang pendidikan geografis. Oleh karena itu, GeoKG mulai

mendefinisikan elemen dasar dan model yang dikonsep dengan menggunakan enam pertanyaan inti dalam geografi. Setiap pertanyaan berhubungan dengan satu masalah inti:

- Dimana itu? →ruang
- Seperti apa itu? →negara bagian
- Mengapa hal itu ada di sana? →evolusi
- Kapan dan bagaimana hal itu terjadi? →perubahan
- Apa dampaknya? →interaksi
- Bagaimana cara mengelolanya demi keuntungan bersama bagi umat manusia dan lingkungan alam? → penggunaan

Elemen Utama

Pertanyaan-pertanyaan di atas dapat digunakan untuk menggambarkan enam aspek inti pengetahuan geografis yang harus diwakili oleh GeoKG. Setiap aspek memerlukan beberapa elemen untuk mendeskripsikannya, dan kami mencoba menemukan elemen dasar di antara semua aspek:

- Ruang →{objek, lokasi, waktu, relasi, ... }
- Status →{objek, waktu, lokasi, atribut ... }
- Evolusi →{objek, keadaan, perubahan, waktu, lokasi, atribut, ... }
- Ubah →{objek, waktu, lokasi, atribut, relasi, ... }
- Interaksi →{objek, relasi, perubahan, ... }
- Penggunaan →{objek, perubahan, status, ... }

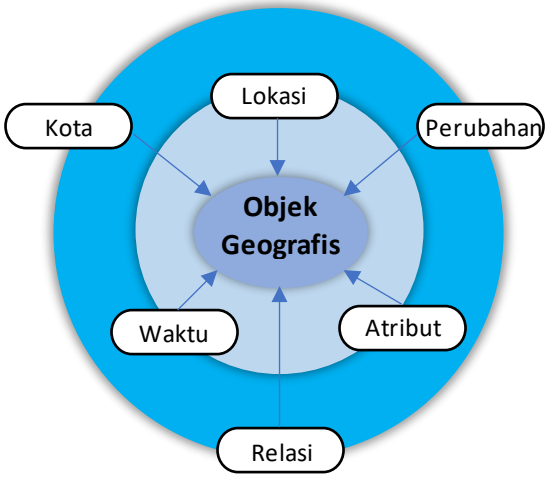
Terdapat tujuh jenis elemen di antara gambaran enam aspek tersebut: objek, lokasi, waktu, atribut, relasi, keadaan, dan perubahan. Tiga ciri khas ketujuh unsur tersebut dalam menggambarkan keenam aspek tersebut adalah sebagai berikut:

- Representasi yang berpusat pada objek. Keseluruhan uraian keenam aspek tersebut memerlukan objek geografis. Tanpa objek, elemen lain tidak ada artinya. Oleh karena itu, enam elemen dasar terbentuk di sekitar elemen objek.
- Representasi gabungan. Deskripsi dari satu elemen dasar hanyalah sebuah pernyataan. Untuk mewakili aspek-aspek ini dalam geografi, elemen-elemen dasar harus digabungkan. Dengan demikian, seluruh elemen dasar dapat dihubungkan.
- Representasi bertahap. Perhatikan bahwa keenam aspek dari pertanyaan inti geografis tidaklah sama. Ruang dan keadaan fokus pada kondisi statis benda. Evolusi dan perubahan lebih memperhatikan kondisi dinamis benda. Selain itu, interaksi dan penggunaan bergantung pada hubungan dan mekanisme antar objek geografis. Oleh karena itu, unsur-unsur dasar tidak dapat diperlakukan secara setara.

Berdasarkan tiga ciri khas elemen dasar tersebut, kami menemukan bahwa objek geografis adalah jenis media yang digunakan untuk merepresentasikan pengetahuan geografis. Ada enam elemen dasar yang digunakan untuk menggambarkan pengetahuan geografis (lihat Gambar 10.2). Lokasi, waktu, atribut, keadaan, perubahan, dan relasi dapat secara efisien mewakili objek geografis dari berbagai aspek. Perhatikan bahwa unsur-unsur dasar ini tidak setara. Lokasi, waktu, dan atribut termasuk dalam tingkat pertama dan mewakili keadaan

statis tunggal suatu objek geografis. Keadaan, perubahan, dan hubungan menggambarkan evolusi dinamis dan hubungan dengan objek geografis.

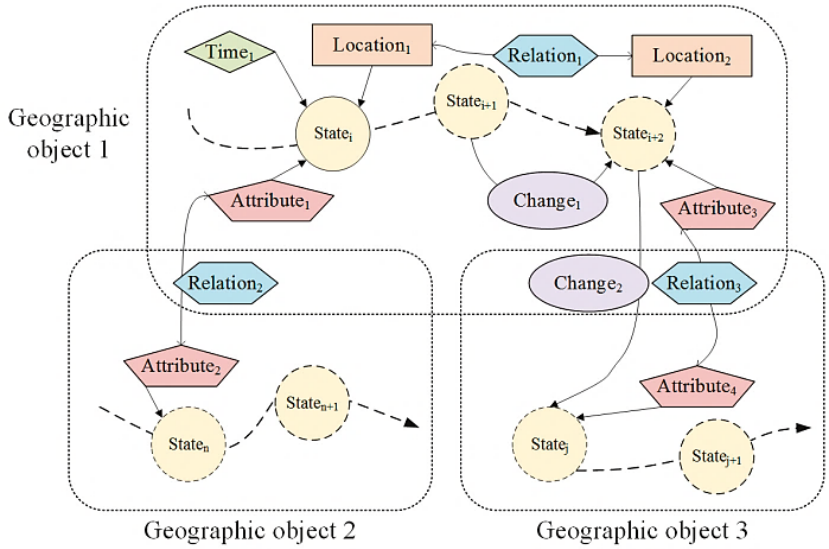
- Objek geografis adalah inti dari representasi pengetahuan geografis dan merupakan unit minimum untuk melihat dunia. Enam elemen dasar (lokasi, waktu, atribut, keadaan, perubahan dan hubungan) mewakili pengetahuan geografis dari perspektif berbeda, yang terkait dengan objek geografis.
- Objek geografis yang independen dan statis dapat dideskripsikan berdasarkan elemen lokasi, waktu, dan atribut. Lokasi menunjukkan pola spasial objek geografis. Waktu memberikan dimensi temporal objek geografis untuk kognisi manusia. Atribut menggambarkan fitur statis objek geografis.
- Setiap objek geografis mempunyai seluruh siklus hidup, termasuk tahapan generasi, perubahan, evolusi, dan kepunahan. Tahapan yang berbeda dalam siklus hidup mewakili keadaan yang berbeda. Negara diwakili oleh kumpulan atribut objek geografis dalam dimensi spasial-temporal tertentu.
- Objek geografis tidak selalu statis. Setiap perubahan pada elemen lain dari suatu objek geografis akan mengubah suatu keadaan menjadi keadaan lain atau suatu relasi ke relasi lain. Dengan demikian, perubahan merupakan bagian penting dari representasi pengetahuan geografis.
- Objek geografis tidak terisolasi. Setiap pemandangan, fenomena, dan lingkungan terdiri dari banyak objek geografis dan hubungan kompleks di antara mereka. Dengan demikian, relasi merupakan deskripsi kunci dari interaksi antar objek geografis yang kompleks.



Gambar 10.2. Enam elemen dasar untuk mewakili suatu objek geografis.

Model GeoKG

Model konseptual GeoKG ditunjukkan pada Gambar 10.3, yang didasarkan pada gagasan yang disebutkan di atas.



Gambar 10.3. Model GeoKG yang dikonsep berdasarkan enam elemen dasar.

Enam elemen inti mewakili objek geografis dan informasinya secara bersamaan. Dalam model ini, objek geografis terdiri dari serangkaian negara. Setiap keadaan suatu objek geografis diwakili oleh atribut-atribut dalam kondisi spasial-temporal tertentu. Dua keadaan yang berkesinambungan atau keadaan yang berbeda antara dua objek geografis dapat menghasilkan elemen perubahan. Unsur perubahan dapat dikategorikan menjadi perubahan waktu, perubahan lokasi atau perubahan atribut. Jika atribut esensial diubah, objek geografis tersebut akan menjadi objek geografis lainnya. Elemen relasi ada antara waktu, lokasi, dan atribut keadaan yang berbeda, terlepas dari apakah mereka merupakan objek geografis yang sama atau tidak.

10.5 FORMALISASI MODEL

Untuk mengatur pengetahuan geografis dengan mempertimbangkan ide-ide dasar, GeoKG harus didasarkan pada model yang menyeluruh dan formal. Bagian ini memberikan model semantik GeoKG dengan menggunakan logika deskripsi (DL), yang tidak terbatas hanya pada tingkat pelengkap bahasa atribut (ALC). Dengan menggunakan logika deskripsi, pengguna dapat membuat deskripsi konseptual untuk representasi dan komputasi pengetahuan geografis yang jelas dan formal.

DL dan Operator Konstruksi

DL terdiri dari tiga komponen dasar: konsep, individu (contoh), dan peran. Konsep mendeskripsikan ciri-ciri umum kumpulan individu, misalnya seluruh daratan yang menonjol jauh di atas lingkungannya membentuk konsep “gunung”. Individu adalah contoh konsep, misalnya entitas geografis, seperti “Pegunungan Rocky”. Peran dapat dijelaskan sebagai hubungan biner antar individu sebagai properti, misalnya hubungan spasial (konjungsi, disjungsi). Sistem logika deskripsi berisi empat bagian. Bagian-bagian ini mencakup seperangkat konstruksi, yang mewakili konsep dan peran, pernyataan tentang terminologi konsep (kotak terminologi, Tbox), pernyataan tentang individu (kotak pernyataan, Abox) dan mekanisme penalaran Tbox dan Abox. Tbox merupakan himpunan yang memuat definisi

hubungan konsep dan aksioma hubungan, yang berisi penjelasan konsep dan peranan. Abox menyertakan kumpulan aksioma yang menggambarkan situasi tertentu, yang berisi informasi contoh Tbox. Abox menyertakan dua bentuk. Salah satunya adalah penegasan konsep, yang mengungkapkan apakah suatu objek termasuk dalam suatu konsep. Yang lainnya adalah penegasan relasi, yang menyatakan apakah dua objek memenuhi relasi tertentu. Logika deskripsi dapat merepresentasikan konsep dan relasi rumit pada konsep atom dan relasi atom berdasarkan operator konstruksi yang diberikan. Operator konstruksi dasar adalah dan (⊓), atau (⊔), bukan (¬), bilangan eksistensial (∃), dan bilangan universal (∀), yang termasuk dalam ALC DL. Lebih banyak operator dapat mewakili lebih banyak logika, yang membentuk jenis DL yang berbeda.

Biarkan C dan D menjadi konsep; a, b, dan c, orang perseorangan; dan R adalah peran antar individu. S adalah peran sederhana, dan n adalah bilangan bulat non-negatif. Seperti biasa, penafsiran $I = (\Delta^I, \cdot^I)$ terdiri dari himpunan tidak kosong Δ^I, yang disebut domain dari I, dan penilaian $\cdot^I$, yang mengasosiasikan, dengan setiap peran R, relasi biner $R^I \subseteq \Delta^I \times \Delta^I$ Untuk bacaan latar belakang yang komprehensif, silakan merujuk ke referensi kertas. Operator utama yang berbeda dari DL ditunjukkan pada Tabel 10.1.

Diagram disediakan untuk mengilustrasikan makna grafis dari operator yang terkait dengan objek geografis dan hubungannya. Konsep teratas menunjukkan semua konsep atau objek, misalnya ⊤ Sungai berarti semua sungai. Konsep terbawah menunjukkan tidak ada konsep atau objek dalam himpunan, misalnya ⊥ Sungai berarti tidak ada sungai dalam himpunan. Konsep atom menunjukkan konsep minimal, misalnya Ac dapat berupa sungai, laut, kota, atau negara. Peran atom menunjukkan hubungan antara dua konsep atom, misalnya R ⊆ sungai × lautan berarti terdapat hubungan antara sungai dan lautan. Konjungsi menunjukkan dua individu yang bersatu atau terhubung, misalnya Sungai Yangzi Nanjing menunjukkan bagian gabungan dari Sungai Yangzi dan Nanjing.

Tabel 10.1. Sintaks dan semantik operator konstruksi utama logika deskripsi.

Category (Symbol)	Construction Operators	Syntax	Semantics	Diagrams
ALC	Top concept	$\top$	Δ^I	
	Bottom concept	$\bot$	$\varnothing$	
	Atomic concept	Ac	$Ac^I \subseteq \Delta^I$	
	Atomic role	R	$R^I \subseteq \Delta^I \times \Delta^I$	
	Conjunction	C⊓D	$C_1^I \cap C_2^I$	
	Disjunction	C⊔D	$C_1^I \cup C_2^I$	
	Negation	¬C	$\Delta^I \backslash C^I$	
	Exist restriction	∃R.C	$\{a \in C^I \mid \exists y, (a,b) \in R^I \,\&\, b \in C^I\}$	

Category (Symbol)	Construction Operators	Syntax	Semantics	Diagrams
ALC	Value restriction	∀R.C	$\{a \in C^I \mid \forall y, (a,b) \in R^I \,\&\, b \in C^I\}$	
H	Concept inclusion	$C_1 \sqsubseteq C_2$	$C_1^I \subseteq C_2^I$	
	Role inclusion	$R \sqsubseteq S$	$R^I \subseteq S^I$	
I	Inverse role	R^-	$\{(a,b) \in R^I \mid (b,a) \in R^I\}$	
R^+	Trans role	Trans(R)	$\{(a,c) \in R^I \mid \exists (a,b) \in R^I \wedge (b,c) \in R^I\}$	
Q	Qualifying at least restriction	R.C ≥ n	$\{a \in C^I \mid \#(\{b \mid (a,b) \in R^I \,\&\, b \in C^I\}) \geq n\}$	
	Qualifying at most restriction	R.C ≤ n	$\{a \in C^I \mid \#(\{b \mid (a,b) \in R^I \,\&\, b \in C^I\}) \leq n\}$	

Disjungsi menunjukkan disjungsi logika dua individu, misalnya Nanjing Sungai Yangzi berarti himpunan kombinasi Sungai Yangzi dan Nanjing. Negasi menunjukkan himpunan semua individu yang tidak termasuk dalam individu target, misalnya ¬ Sungai Yangzi berarti semua individu kecuali Sungai Yangzi. Pembatasan yang ada menunjukkan keberadaan individu atau peran, misalnya, ∃ Sungai Yangzi berarti terdapat Sungai Yangzi dan ∃R ⊆ Sungai Yangzi × Pegunungan Zhong berarti terdapat peran antara Sungai Yangzi dan Pegunungan Zhong. Pembatasan nilai menunjukkan semua individu atau peran, misalnya ∀ Sungai berarti semua sungai dan ∀ R ⊆ Sungai Yangzi × Pegunungan Zhong berarti semua peran antara Sungai Yangzi dan Pegunungan Zhong. Inklusi konsep menunjukkan suatu konsep termasuk dalam konsep lain, misalnya hujan ± presipitasi artinya hujan adalah sejenis presipitasi. Penyertaan peran menunjukkan peran yang termasuk dalam rangkaian peran, misalnya Rlokasi_Sungai Yangzi–Gunung Zhong ± Sungai Yangzi × Pegunungan Zhong menunjukkan bahwa hubungan lokasi antara Sungai Yangzi dan Pegunungan Zhong adalah salah satu peran dari keseluruhan rangkaian peran Yangzi Sungai dan Pegunungan Zhong. Peran terbalik menunjukkan bahwa suatu peran memiliki reversibilitas. Peran trans menunjukkan bahwa suatu peran memiliki transmisibilitas. Pembatasan yang memenuhi syarat setidaknya/paling banyak menunjukkan adanya setidaknya atau paling banyak, misalnya (∃≥ 3 sungai) ⊆ Sungai Yangzi berarti Sungai Yangzi memiliki setidaknya tiga cabang.

Representasi Formalisasi

Pada bagian ini, semantik model GeoKG didefinisikan. Pertama, kita tentukan himpunan pengetahuan geografis GK yang bersumber dari seluruh fenomena alam dan manusia di dunia W. GeoGK merupakan himpunan GK yang dapat didefinisikan sebagai berikut:

$$\mathrm{GeoGK} = \{\langle \mathrm{GK} \rangle | \mathrm{GK} \in \mathrm{W}\}$$

GK adalah tupel yang terdiri dari objek geografis O dan elemen dasarnya E:

$$\mathrm{GK} = \{\langle \mathrm{O}, E \rangle | \exists \mathrm{O} \neq \emptyset, \exists E \neq \emptyset\}$$

Kumpulan elemen dasar E berisi enam elemen berbeda: lokasi L, waktu T, atribut A, status St, perubahan Ch, dan relasi Re. Jadi, E adalah tupel enam:

$$\mathrm{E} = \{\langle L, T, A, St, Ch, Re \rangle | \exists L \parallel T \parallel A \parallel St \parallel Ch \parallel Re \neq \emptyset\}$$

Setiap elemen diidentifikasi sebagai berikut:

(1) Waktu

Waktu menggambarkan informasi temporal keadaan suatu objek geografis. Misalkan Sti menunjukkan keadaan spesifik dari objek geografis O_i; elemen dasar waktu T dapat didefinisikan sebagai berikut:

$$\mathrm{T} = \{\exists L \in St \mid \forall O_i \neq \emptyset, St_i \neq O_i\}$$

Waktu harus dijelaskan dengan tipe dasar dan informasi waktu referensi. Tipe dasarnya adalah waktu titik, waktu interval, dan waktu referensi. Waktu titik T_{poi} mencatat

momen keadaan suatu objek geografis. Waktu interval Warna menunjukkan interval waktu antara dua waktu titik. Waktu referensi T_{ref} menunjukkan waktu elemen lain dari suatu objek geografis, misalnya, "Piala Dunia 2018" adalah peristiwa dengan periode waktu unik yang dapat merujuk waktu tertentu secara akurat. Pengetahuan referensi waktu $tref$ menunjukkan tambahan pengetahuan tentang deskripsi waktu. Misalkan tw menunjukkan kata waktu. Kata waktu menunjukkan suatu titik waktu yang dapat berisi beberapa bagian deskriptif waktu, misalnya, 12-Juli-2018, pukul sembilan lewat sepuluh, dan besok pagi. Waktu titik T_{poi}, waktu interval T_{int} dan waktu referensi T_{ref} didefinisikan sebagai berikut:

$$T_{poi} = \{\langle tw, tref\rangle | \forall! \, tw \in T\}$$
$$T_{int} = \{\langle tw, tref\rangle | \forall tw \in T, \#tw \geq 2, \forall R \sqsubseteq tw\}$$
$$T_{ref} = \{\langle E, tref\rangle | \forall E \; \& \; \forall! \, T \sqsubseteq St_i\}$$

dimana R adalah hubungan interval dua kata waktu. Pengetahuan referensi waktu $tref$ merupakan himpunan pengetahuan referensi yang terdiri dari kesamaan, relativitas, ketidakjelasan, kontinuitas, dan periodisitas yaitu, $tref = \{\langle com, rel, fuz, con, per\rangle\}$. Ada beberapa contoh untuk setiap kata referensi waktu. Misalnya, "12-Juli-2018" adalah waktu yang umum, dan Jurassic Akhir adalah deskripsi waktu domain. Relativitas menunjukkan apakah waktu itu relatif, misalnya, "dua hari yang lalu" adalah waktu relatif yang mengacu pada waktu absolut "hari ini", "jam 9" adalah waktu yang akurat, "sekitar jam 9" adalah waktu fuzzy, "12 -Juli" adalah waktu instan, dan "hingga 12 Juli" adalah waktu berkelanjutan. Periodisitas dapat dengan mudah dipahami, seperti "setiap akhir pekan", "setiap bulan", dan "setiap tahun".

(2) Lokasi

Lokasi menggambarkan informasi spasial keadaan suatu objek geografis. Misalkan St_i menunjukkan keadaan spesifik dari objek geografis O_i; lokasi elemen dasar L dapat diidentifikasi sebagai berikut:

$$L = \{\exists L \in St_i | \forall O_i \neq \emptyset, St_i \in O_i\}$$

Menurut kompleksitas deskripsi lokasi, suatu lokasi dapat diatur menjadi tipe dasar dan informasi referensi lokasi. Tipe dasarnya meliputi toponim, alamat, koordinat, dan lokasi referensi. Toponim L_{top} menggambarkan suatu lokasi dengan nama umum. Alamat L_{add} menunjukkan lokasi dengan nomor teratur dan jalan yang diberi nama oleh administrator. Koordinat L_{coo} mencatat lokasi dengan serangkaian angka yang disusun secara matematis. Lokasi referensi L_{ref} menunjukkan lokasi elemen lain dalam suatu objek geografis. Pengetahuan referensi lokasi menunjukkan pengetahuan tambahan tentang deskripsi lokasi. Misalkan tp, ad, co masing-masing menunjukkan toponim, alamat, dan koordinat. Toponim L_{top}, alamat L_{add}, koordinat L_{coo}, dan lokasi referensi L_{ref} diidentifikasi sebagai berikut:

$$L_{top} = \{\langle tp, lref\rangle | \forall! \, tp \in L\}$$
$$L_{add} = \{\langle ad, lref\rangle | \forall! \, tp \in L\}$$

$$L_{coo} = \{\langle co, lref\rangle | \forall! \, tp \in L\}$$
$$L_{ref} = \{\langle E, lref\rangle | \forall E \; \& \; \forall! L \sqsubseteq St_i\}$$

Pengetahuan referensi lokasi $lref$ merupakan himpunan pengetahuan referensi yang terdiri dari tipe ruang, referensi spasial, kesamaan, relativitas, dan ketidakjelasan yaitu, $lref = \{\langle typ, ref, com, rel, fuz\rangle\}$. Tipe ruang menjelaskan tipe ruang apa, seperti realitas, virtual, atau lokasi domain tertentu. Misalnya, Pandora adalah toponim dari dunia maya film Avatar. Referensi spasial menggambarkan sistem deskripsi lokasi, misalnya proyeksi WGS84 dan Mercator. Kesamaan menyimpan apakah suatu lokasi merupakan lokasi domain, misalnya, Beijing adalah toponim umum yang dapat dikodekan sebagai "-.-..--...-.---/-..---.-.-. --.." dalam sistem kode Morse. Relativitas menunjukkan apakah suatu lokasi bersifat relatif, misalnya, "20 km selatan Beijing" adalah deskripsi lokasi relatif yang mengacu pada lokasi absolut "Beijing". Ketidakjelasan menyatakan apakah deskripsi lokasi akurat atau tidak, misalnya "dekat Times Square" adalah deskripsi lokasi yang kabur.

(3) Atribut

Atribut menggambarkan informasi fitur keadaan suatu objek geografis. Misalkan St_i menunjukkan keadaan spesifik dari objek geografis O_i; atribut elemen dasar A dapat diidentifikasi sebagai berikut:

$$A = \{\exists A \in St_i | \forall O_i \neq \emptyset, St_i \in O_i\}$$

Semua deskripsi fitur suatu objek geografis termasuk dalam suatu atribut, misalnya bentuk, warna, kecepatan, dll. Untuk mengatur atribut suatu objek geografis, mengidentifikasi apa yang dimaksud dengan atribut adalah kuncinya. Atribut adalah deskripsi fitur tunggal dari suatu objek geografis. Misalnya, "topan adalah siklon tropis matang yang berkembang antara 180◦ dan 100◦ BT di Belahan Bumi Utara, dengan puncak bulan Agustus hingga Oktober" menjelaskan tiga atribut: atribut tipikal "siklon tropis matang", atribut lokasi dari "berkembang antara 180◦ dan 100◦ BT di Belahan Bumi Utara" dan atribut frekuensi "bulan puncak dari Agustus hingga Oktober". Perlu diketahui bahwa atribut dapat dibagi menjadi dua jenis: atribut esensial A_{es} dan atribut non-esensial A_{ne}:

$$A_{es} = \{\exists A_{es} \in St_i | \forall O_i \neq \emptyset, St_i \in O_i, \#A_{es} \geq 1\}$$
$$A_{ne} = \{A_{ne} \in A' | \forall O_i \neq \emptyset, St_i \in O_i, \#A' = {A_{es}}^{-}\}$$

Atribut esensial adalah atribut tanda yang mengidentifikasi suatu objek geografis dari yang lain. Ketika atribut penting berubah, suatu objek geografis dapat berubah menjadi objek lain. Misalnya, ketika siklon tropis dewasa berkembang di Samudera Atlantik, hal tersebut tidak mungkin merupakan topan. Atribut non-esensial adalah deskripsi fitur lain dari suatu objek geografis, misalnya atribut frekuensi angin topan adalah "bulan puncak dari Agustus hingga Oktober". Atribut-atribut ini tidak dapat menentukan sifat suatu objek geografis.

(4) Negara

Negara menggambarkan tahapan yang berbeda dari suatu objek geografis. Terlihat bahwa ketiga unsur dasar di atas bekerja sama untuk menyatakan keadaan. Dengan demikian, keadaan unsur St dapat diidentifikasi sebagai berikut:

$$St = \{\exists St_i \in O | \exists! \, L \sqsubseteq St_i, \exists! \, T \sqsubseteq St_i, \exists A \sqsubseteq St_i, A \geq 0\}$$

di mana ∃! berarti keberadaan yang unik. Rumusannya mengandung arti bahwa negara merupakan bagian dari suatu objek geografis. Karena keadaan elemen St diwakili oleh kumpulan atribut objek geografis dalam dimensi spasial-temporal tertentu, maka elemen tersebut harus bergantung pada lokasi elemen L dan elemen Waktu. Perhatikan bahwa elemen lokasi L dan elemen Waktu T ada secara unik, karena waktu dan ruang adalah dua dimensi yang mewakili tahapan dalam ruang Euclidean. Misalnya, keadaan topan mencakup semua fitur untuk kerangka referensi spasial-temporal tertentu, misalnya, "Topan Maria, 23:00/10Juli-2018, E123.40◦/N25.60◦, tekanan pusat 945 hpa, maks kecepatan 30 km/jam". Negara tidak dapat didefinisikan tanpa informasi temporal dan spasial. Sebaliknya, keadaan elemen St tidak bergantung pada elemen Atribut A. Atribut adalah catatan deskriptif yang tidak dapat mempengaruhi apakah keadaan itu ada. Misalnya, "Topan Maria, 23:00/10Juli-2018, E123.40◦/N25.60◦" juga mendefinisikan keadaan Topan Maria. Dengan demikian, elemen atribut didefinisikan berbeda dengan elemen lokasi dan elemen waktu.

(5) Perubahan

Perubahan menggambarkan perubahan suatu objek geografis dari satu keadaan ke keadaan lainnya. Dengan demikian, perubahan Ch harus mengandung setidaknya satu perbedaan antara dua keadaan, yang dapat berupa perubahan lokasi, perubahan waktu, atau perubahan atribut. Perubahan mengandung empat komponen utama:

$$Ch = \{\langle St, act, CE, type\rangle \in O | \exists St, \#St = 2, CE \in \{T, L, A\}, type \in (Ch_d, Ch_e)\}$$

dimana St menunjukkan keadaan (termasuk dua keadaan yang berbeda), tindakan menunjukkan tindakan perubahan, CE menunjukkan elemen perubahan dan tipe menunjukkan jenis perubahan. Perlu dicatat bahwa ada dua jenis perubahan: perubahan yang berkembang dan perubahan yang berevolusi. Perubahan yang berkembang menunjukkan perubahan dari satu objek geografis, dan perubahan yang berkembang menggambarkan perubahan antara dua objek geografis yang berbeda. Misalkan Ch_d menunjukkan perubahan yang berkembang dan Ch_e menunjukkan perubahan yang berkembang; definisi formalnya adalah sebagai berikut:

$$Ch_d = \{\exists Ch_d = St_i \times St_{i+1} | \exists St_i \& St_{i+1} \in O_m, St_i \neq St_{i+1}\}$$
$$Ch_e = \{\exists Ch_e = St_{end} \times St_i | \exists St_{end} \in O_m, \exists St_i \in O_n, \exists St_{end} . A_{es} \neq St_i . A_{es}\}$$

dimana $O, O_m,$ dan O_n adalah objek geografis, St_i dan St_{i+1} menunjukkan keadaan berkelanjutan dari objek geografis, Stend menunjukkan keadaan terakhir dari objek geografis, dan A_{es} menunjukkan atribut penting dari objek geografis.

(6) Hubungan

Suatu relasi mengungkapkan perbedaan antara elemen objek geografis, yang mencakup tiga tipe umum: relasi lokasi, relasi waktu, dan relasi atribut. Ketiga jenis ini masing-masing menggambarkan perbedaan spasial, perbedaan waktu, dan perbedaan fitur. Suatu relasi mengandung tiga komponen utama: elemen dari dua keadaan E, semantik relasi Sem, dan tipe $type$ relasi:

$$Re = \{\langle E, Sem, type\rangle \in O | \exists E \&\# E \geq 2, type \in (Re_l, Re_t, Re_a)\}$$

Misal Re_l, Re_t, dan Re_a masing-masing menunjukkan hubungan lokasi, hubungan waktu, dan hubungan atribut, L_i dan L_j menunjukkan lokasi keadaan yang berbeda, T_i dan T_j menunjukkan waktu keadaan yang berbeda, dan A_i dan $A_j$menunjukkan atribut keadaan yang berbeda. Berbagai jenis hubungan diidentifikasi sebagai berikut:

$$Re_l = \{\exists Re_l = L_i \times L_j | \exists St_i \& St_j, St_i \neq St_{i+1}\}$$
$$Re_t = \{\exists Re_t = L_i \times L_j | \exists St_i \& St_j, St_i \neq St_{i+1}\}$$
$$Re_a = \{\exists Re_a = L_i \times L_j | \exists St_i \& St_j, St_i \neq St_{i+1}\}$$

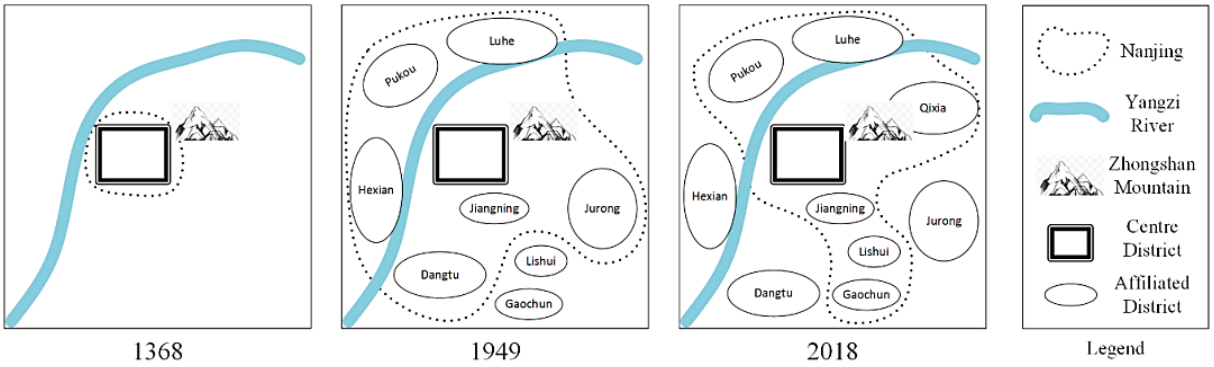
Relasi lokasi menggambarkan hubungan spasial antara negara bagian yang berbeda, misalnya hubungan lokasi antara negara bagian yang berbeda saat terjadi topan atau hubungan lokasi antara dua pusat kota berbeda yang sedang dikembangkan. Relasi waktu menggambarkan hubungan temporal antara negara-negara yang berbeda, yaitu rentang waktu antara dua negara, misalnya rentang waktu pengalihan sungai. Relasi atribut menggambarkan hubungan fitur antara keadaan yang berbeda, yaitu perbedaan antara dua keadaan topan, misalnya kecepatan angin maksimal, tekanan pusat, dll.

10.6 STUDI KASUS

Pada bagian ini, contoh lengkap ditampilkan untuk mengilustrasikan representasi pengetahuan geografis menggunakan model GeoKG. Untuk menggambarkan representasi pengetahuan geografis dengan jelas, kasus evolusi pembagian administratif Nanjing dipilih. Contoh yang diberikan mencakup objek geografis dasar (misalnya, Sungai Yangzi, Gunung Zhongshan), wilayah Nanjing yang berubah, dan beberapa distrik yang berafiliasi di era yang berbeda.

Area penelitian

Nanjing, sebelumnya diromanisasi menjadi Nanking dan Nankin, adalah ibu kota provinsi Jiangsu di Republik Rakyat Tiongkok dan kota terbesar kedua di wilayah Tiongkok Timur, dengan wilayah administratif lebih dari 6000 km2. Wilayah dalam Nanjing yang dikelilingi tembok kota adalah Distrik Pusat Nanjing, dengan luas 55 km2, sedangkan Wilayah Metropolitan Nanjing meliputi kota dan wilayah sekitarnya. Tiga tahapan representatif dipilih untuk mewakili revolusi Nanjing: 1368, 1949, dan 2018. Sketsa peta ditunjukkan pada Gambar 10.4.



Gambar 10.4. Peta sketsa evolusi pembagian administratif Nanjing pada tahun 1368, 1949, dan 2018.

Tahap pertama adalah Dinasti Ming, yang pertama kali menamai kota ini dengan kata "Nanjing". Kaisar pertama dinasti Ming, Zhu Yuanzhang, yang menggulingkan dinasti Yuan, mengganti nama kota Nanjing, membangunnya kembali, dan menjadikannya ibu kota dinasti pada tahun 1368. Ia membangun tembok kota sepanjang 48 km di sekitar Nanjing. Itu adalah distrik pusat Nanjing, yang terletak di selatan Sungai Yangzi dan di sebelah barat Gunung Zhongshan.

Tahap kedua adalah berdirinya Republik Rakyat Tiongkok. Pemerintah menetapkan Nanjing sebagai satuan provinsi yang dikontrol langsung oleh pemerintah. Pada tahap itu, Nanjing mengelola distrik pusat dan beberapa distrik afiliasinya. Distrik pusat mencakup distrik 1–10 dan distrik afiliasinya mencakup Jiangning, Jurong, Dangtu, Hexian, Pukou, dan Luhe. Pada tahun 1949, Nanjing telah disalurkan melalui Sungai Yangzi dan Gunung Zhongshan.

Tahap ketiga adalah tahun 2018 yang mengacu pada batas administratif Nanjing saat ini. Setelah serangkaian penyesuaian pembagian administratif, Gaochun dan Lishui dimasukkan ke dalam Nanjing dan Jurong, Dangtu, dan Hexian dipindahkan dari perbatasan.

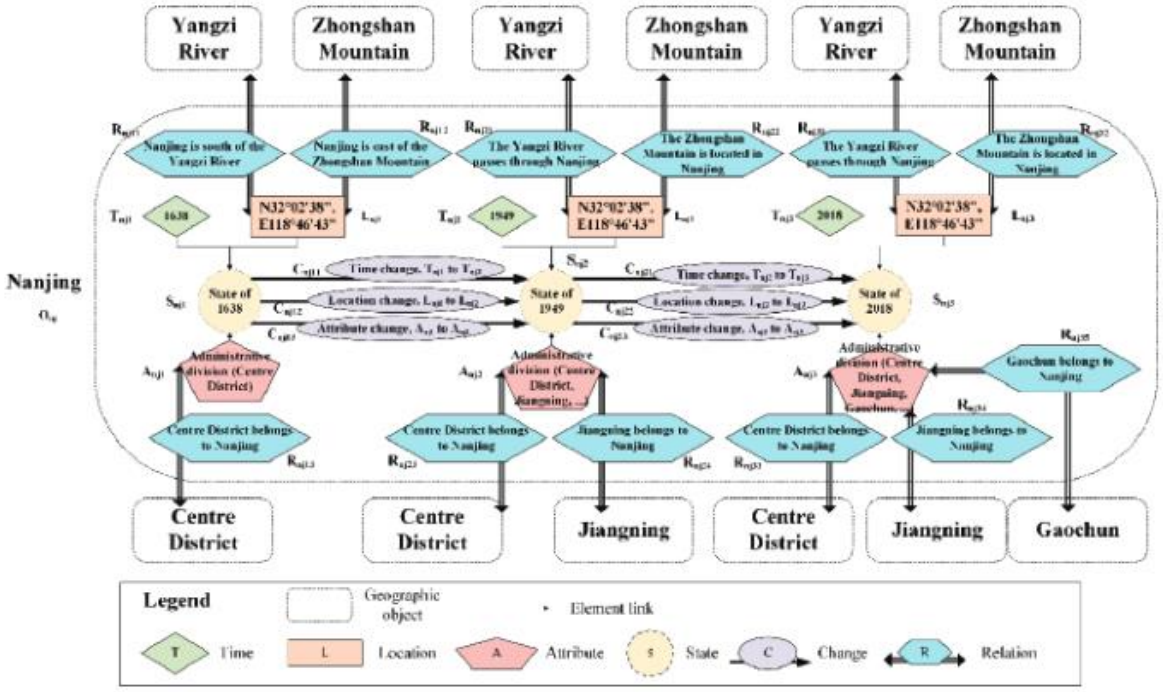
Selama lebih dari 600 tahun pembangunan Nanjing, banyak elemen diubah termasuk batas, distrik yang berafiliasi, hubungan antara Nanjing dan objek geografis lainnya (misalnya, Sungai Yangzi dan Gunung Zhongshan). Hubungan yang berbeda terjadi pada tahapan yang berbeda di antara objek-objek geografis ini. Oleh karena itu, model GeoKG digunakan untuk mewakili perubahan pengetahuan geografis ini. Formalisasi diperkenalkan pada bagian berikutnya.

Formalisasi

Dalam contoh ini, evolusi pembagian administratif diatur dengan menggunakan model GeoKG. Objek geografis adalah kunci untuk merepresentasikan pengetahuan geografis. Pertama, kasus ini mengidentifikasi enam objek geografis yang relevan: Nanjing O_{nj}, Sungai Yangzi O_{yr}, Zhongshan Mountain O_{zm}, Center District O_{cd}, Jiangning O_{jn}, dan Gaochun O_{gc}. Jiangning dan Gaochun merupakan distrik afiliasi perwakilan yang dipilih dalam kasus ini. Jiangning selalu menjadi bagian dari Nanjing pada tahun 1949 dan 2018 dan Gaochun mengalami penyesuaian pembagian administratif. Setiap objek geografis terdiri dari serangkaian keadaan, perubahan, dan hubungan. Misalnya, Nanjing O_{nj} berisi tiga keadaan

$S_{nj} = \{S_{nj1}, S_{nj2}, S_{nj3}\}$, enam perubahan $C_{nj} = \{C_{nj11}, C_{nj12}, C_{nj13}, C_{nj21}, C_{nj22}, C_{nj23}\}$ dan 12 relasi $R_{nj} = \{R_{nj11}, R_{nj12}, R_{nj13,} R_{nj21}, R_{nj22}, R_{nj23}, R_{nj24}, R_{nj31}, R_{nj32}, R_{nj34}, R_{nj35}\}$. Dengan demikian, Nanjing O_{nj} dapat didefinisikan sebagai berikut dan diagram terkait ditunjukkan pada Gambar 10.5.

$$O_{nj} = \left\{ \begin{array}{c} S_{nj} \sqsubseteq O_{nj}, C_{nj} \sqsubseteq O_{nj}, R_{nj} \sqsubseteq O_{nj} | \\ S_{nj} = \{S_{nj1}, S_{nj2}, S_{nj3} | S_{nj}.number \leq 3, S_{nj}.number \geq 3\}, \\ C_{nj} = \{C_{nj11}, C_{nj12}, C_{nj13}, C_{nj21}, C_{nj22}, C_{nj23} | C_{nj}.number \leq 6, C_{nj}.number \geq 6\} \\ R_{nj} = \left\{ \begin{array}{c} R_{nj11}, R_{nj12}, R_{nj13}, R_{nj21}, R_{nj22}, R_{nj23}, R_{nj24}, R_{nj31}, R_{nj32}, R_{nj33,} R_{nj34} R_{nj35} | \\ R_{nj}.number \leq 12, R_{nj}.number \geq 12 \end{array} \right\} \end{array} \right\}$$



Gambar 10.5. Diagram berbagai elemen Nanjing dengan menggunakan model GeoKG.

Sebenarnya, negara bagian Nanjing $S_{nj1}, S_{nj2}, S_{nj3}$ yang berbeda menunjukkan tiga tahapan berbeda yaitu tahun 1368, 1949, dan 2018. Setiap negara bagian berisi elemen waktu, lokasi, dan atribut yang berbeda. Misalnya, negara bagian S_{nj1} di Nanjing berisi elemen waktu T_{nj} dari "1368", elemen lokasi L_{nj} dari "deskripsi lokasi pada tahun 1368" dan elemen atribut A_{nj} dari "wilayah administratif". S_{nj1} negara bagian Nanjing dapat didefinisikan sebagai berikut:

$$S_{nj1} = \left\{ \begin{array}{c} T_{nj} \sqsubseteq S_{nj1}, L_{nj} \sqsubseteq S_{nj1}, A_{nj} \sqsubseteq S_{nj1} | \\ T_{nj} = \{T_{nj1} | T_{nj}.number \leq 1, T_{nj}.number \geq 1\}, \\ L_{nj} = \{L_{nj1} | L_{nj}.number \leq 1, L_{nj}.number \geq 1\}, \\ A_{nj} = \{A_{nj1} | A_{nj}.number \leq 1, A_{nj}.number \geq 1\} \end{array} \right\}$$

Negara bagian yang berbeda dapat memuat perubahan yang menunjukkan jenis perubahan yang berbeda dari satu negara bagian ke negara bagian lainnya. Misalnya, ada tiga perubahan utama $\{C_{nj11}, C_{nj12}, C_{nj13}\}$ dari negara bagian S_{nj1} Nanjing pada tahun 1368 menjadi negara bagian S_{nj2} Nanjing pada tahun 1949: perubahan C_{nj11} antar elemen waktu, perubahan C_{nj12} antar elemen lokasi, dan perubahan C_{nj13} antar unsur atribut "wilayah administratif". Perhatikan bahwa semua perubahan ini termasuk dalam jenis perubahan pengembangan yang menunjukkan bahwa perubahan tersebut tidak membuat objek geografis baru. Perubahan tersebut dapat didefinisikan sebagai berikut:

$$C_{nj11} = \begin{Bmatrix} St, act, CE, type \sqsubseteq C_{nj11} | \\ St = \{S_{nj1}, S_{nj2}\}, act = \{"time\ change"\}, CE = \{T_{nj1}, T_{nj2}\}, type = C_{hd} \end{Bmatrix} \sqsubseteq O_{nj}$$

$$C_{nj12} = \begin{Bmatrix} St, act, CE, type \sqsubseteq C_{nj12} | \\ St = \{S_{nj1}, S_{nj2}\}, act = \{"time\ change"\}, CE = \{L_{nj1}, L_{nj2}\}, type = C_{hd} \end{Bmatrix} \sqsubseteq O_{nj}$$

$$C_{nj13} = \begin{Bmatrix} St, act, CE, type \sqsubseteq C_{nj13} | \\ St = \{S_{nj1}, S_{nj2}\}, act = \{"time\ change"\}, CE = \{A_{nj1}, A_{nj2}\}, type = C_{hd} \end{Bmatrix} \sqsubseteq O_{nj}$$

Relasi merupakan elemen yang sangat diperlukan yang ada pada objek geografis yang mengacu pada hubungan antar elemen yang berbeda. Dalam contoh ini, terdapat tiga relasi $R_{nj11}, R_{nj12}, R_{nj13}$ yang berhubungan dengan Nanjing pada tahun 1368: relasi spasial R_{nj11} antara Nanjing O_{nj} dan Sungai Yangzi O_{yz}, relasi spasial R_{nj12} antara Nanjing O_{nj} dan Zhongshan Mountain O_{zm}, dan relasi atribut R_{nj13} antara Nanjing O_{nj} dan Center District O_{cd}, dimana L_{yz1} adalah lokasi Sungai Yangzi O_{yz}, pada tahun 1368, L_{zm1} adalah lokasi Zhongshan Mountain O_{zm} pada tahun 1368 dan A_{cd1} adalah atribut "wilayah administratif" dari Center District O_{cd} pada tahun 1368. Hubungan tersebut dapat didefinisikan sebagai berikut dan diagram hubungan ini ditunjukkan pada Gambar 10.6.

$$R_{nj11} = \begin{Bmatrix} E, Sem, type \sqsubseteq R_{nj11} | \\ E = \{L_{nj1}, L_{yz2}\}, Sem = \{"Nanjing\ is\ south\ of\ the\ Yangzi\ River"\}, type = Re_l \end{Bmatrix} \sqsubseteq O_{nj}$$

$$R_{nj12} = \begin{Bmatrix} E, Sem, type \sqsubseteq R_{nj12} | \\ E = \{L_{nj1}, L_{zm1}\}, Sem = \{"Nanjing\ is\ south\ of\ the\ Yangzi\ River"\}, type = Re_l \end{Bmatrix} \sqsubseteq O_{nj}$$

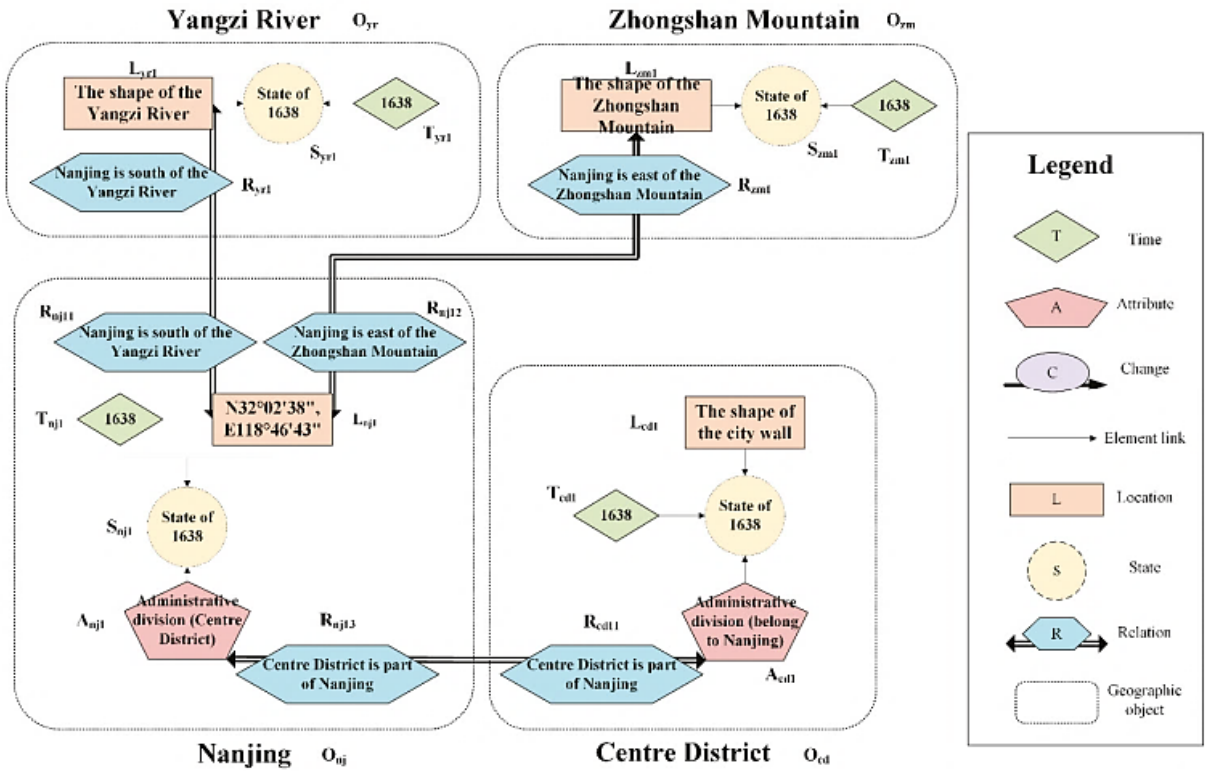
$$R_{nj13} = \begin{Bmatrix} E, Sem, type \sqsubseteq R_{nj13} | \\ E = \{A_{nj1}, A_{cd1}\}, Sem = \{"Nanjing\ is\ south\ of\ the\ Yangzi\ River"\}, type = Re_a \end{Bmatrix} \sqsubseteq O_{nj}$$

Sejalan dengan itu, Sungai Yangzi berisi relasi $R_{yz1} = R_{nj11}^{-}$, Gunung Zhongshan berisi relasi $R_{zm1} = R_{nj12}^{-}$, dan Distrik Tengah berisi relasi $R_{cd1} = R_{nj13}^{-}$,:

$$R_{yz1} = \left\{ \begin{matrix} E, Sem, type \sqsubseteq R_{yz1} | \\ E = \{L_{nj1}, L_{yz1}\}, Sem = \{"Nanjing\ is\ south\ of\ the\ Yangzi\ River"\}, type = Re_l \end{matrix} \right\} \sqsubseteq O_{yz}$$

$$R_{zm1} = \left\{ \begin{matrix} E, Sem, type \sqsubseteq R_{zm1} | \\ E = \{L_{nj1}, L_{zm1}\}, Sem = \{"Nanjing\ is\ south\ of\ the\ Yangzi\ River"\}, type = Re_l \end{matrix} \right\} \sqsubseteq O_{zm}$$

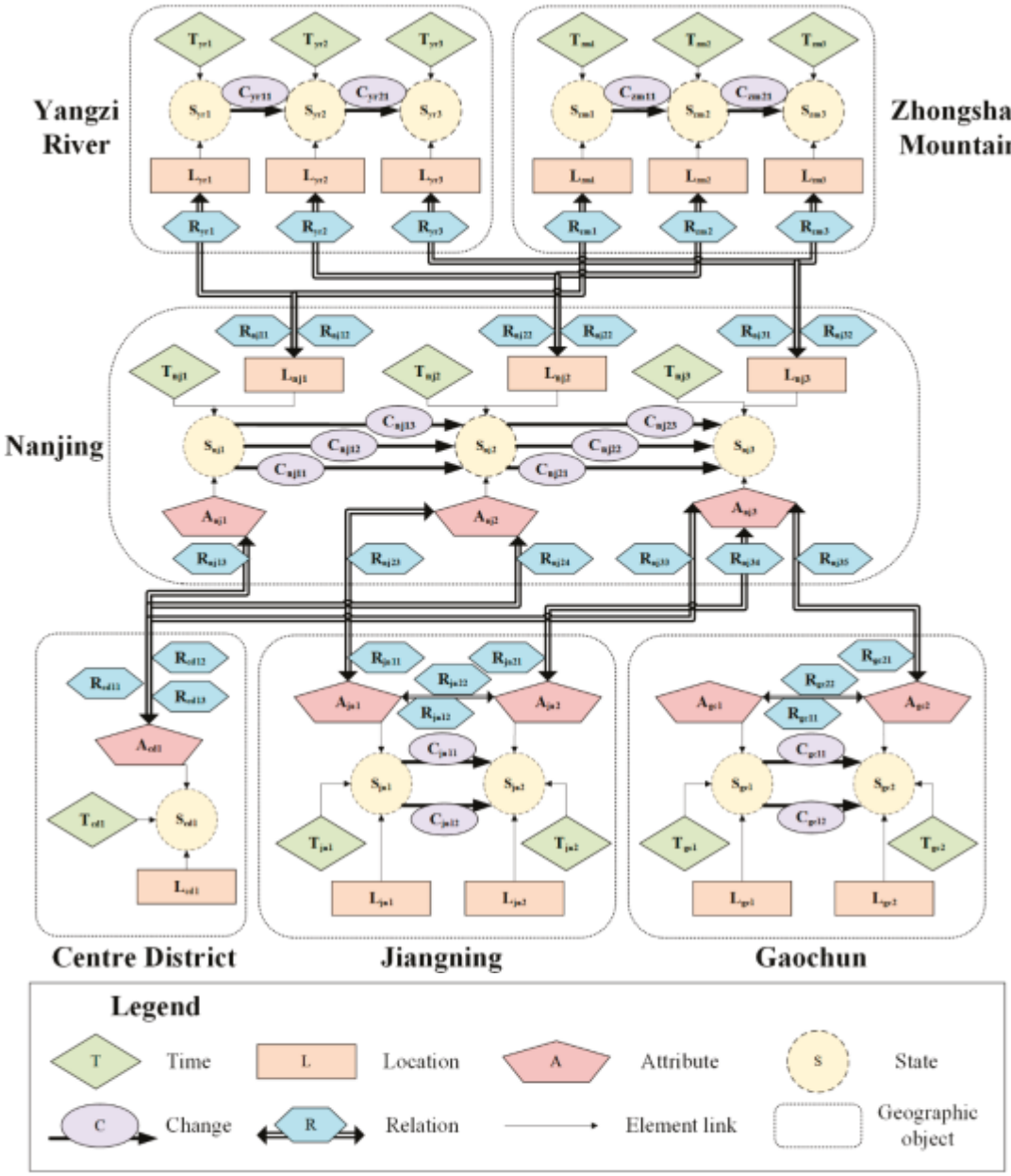
$$R_{cd1} = \left\{ \begin{matrix} E, Sem, type \sqsubseteq R_{cd1} | \\ E = \{A_{nj1}, A_{cd1}\}, Sem = \{"Nanjing\ is\ south\ of\ the\ Yangzi\ River"\}, type = Re_a \end{matrix} \right\} \sqsubseteq O_{cd}$$



Gambar 10.6. Diagram relasi elemen Nanjing tahun 1368.

Keseluruhan kasus evolusi pembagian administratif Nanjing dapat ditunjukkan pada Gambar 10.7. Sesuai dengan Gambar 10.4, setiap objek geografis berisi satu hingga tiga negara bagian. Misalnya, Sungai Yangzi dan Gunung Zhongshan memiliki tiga tahapan pada tahun 1368, 1949, dan 2018, sedangkan Jiangning dan Gaochun memiliki dua tahapan pada tahun 1949 dan 2018. Karena perubahan dalam tidak dipertimbangkan, Distrik Tengah hanya mewakili satu tahapan. Di antara tahapan yang berbeda, jenis perubahan yang berbeda juga dipertimbangkan. Misalnya, tahapan berbeda di Sungai Yangzi dan Gunung Zhongshan mencakup perubahan waktu $\{C_{yz11}, C_{yz21}, C_{zm11}, C_{zm21}\}$ tahapan berbeda di Nanjing mencakup perubahan waktu $\{C_{nj11}, C_{nj21}\}$ perubahan lokasi $\{C_{nj12}, C_{nj22}\}$ dan perubahan atribut $\{C_{nj13}, C_{nj23}\}$ dan tahapan berbeda dari Jiangning dan Gaochun menyertakan perubahan waktu $\{C_{nj11}, C_{gc11}\}$ dan perubahan atribut $\{C_{nj12}, C_{gc12}\}$. Selain itu, relasi menghubungkan elemen-elemen yang berbeda di antara objek geografis yang berbeda dan objek geografis yang sama. Misalnya, Nanjing pada tahun 1368 memiliki hubungan dengan

Sungai Yangzi R_{nj11}, Gunung Zhongshan R_{nj12}, dan Distrik Pusat R_{nj13}. Kemudian, Nanjing pada tahun 1949 memiliki hubungan dengan Sungai Yangzi R_{nj21}, Gunung Zhongshan R_{nj22}, Distrik Pusat R_{nj23}, dan Jiangning R_{nj24}. Pada tahun 2018, Nanjing mempunyai hubungan dengan Sungai Yangzi R_{nj31}, Gunung Zhongshan R_{nj32}, Distrik Pusat R_{nj33}, Jiangning R_{nj34}, dan Gaochun R_{nj35}.



Gambar 10.7. Tinjauan kasus evolusi pembagian administratif Nanjing dan objek geografis terkait.

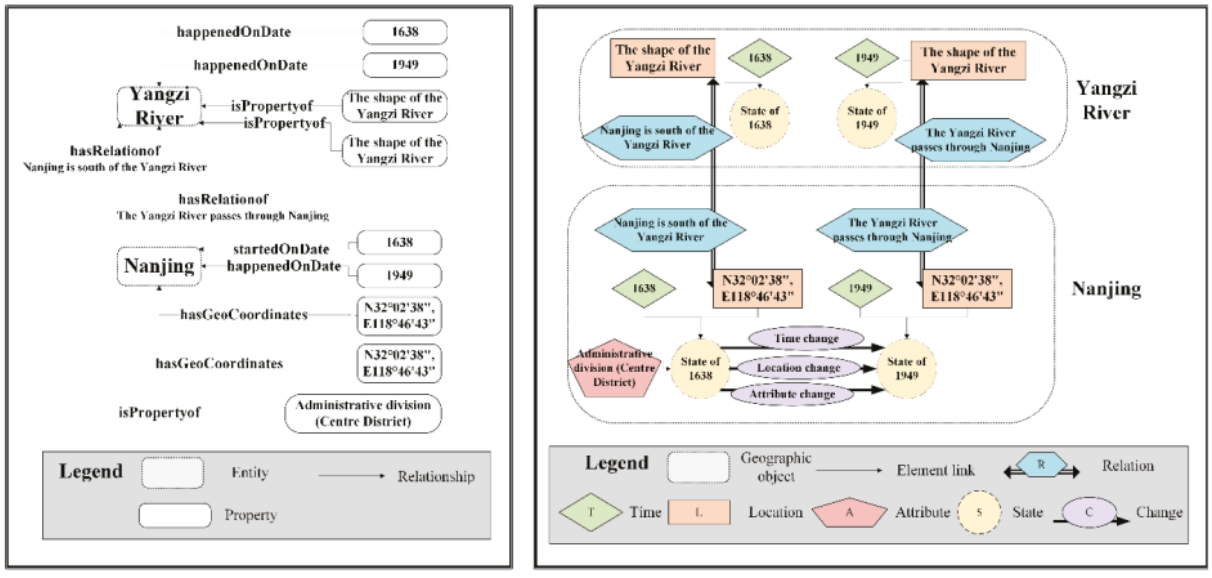
Perhatikan bahwa ada juga hubungan batin antar elemen. Dalam hal ini, pembagian administratif Jiangning pada tahun 1949 memiliki relasi atribut R_{nj12} "hubungan warisan" dengan pembagian administratif Jiangning pada tahun 2018. Gaocun memiliki relasi atribut

yang sama R_{gc11}. Semua hubungan ini mempunyai hubungan terbalik pada arah yang berlawanan.

Pada bagian ini, studi kasus evolusi pembagian administrasi Nanjing dibangun dengan menggunakan model GeoKG dan model YAGO. YAGO adalah grafik pengetahuan sumber terbuka yang representatif dengan versi berbeda. Perhatikan bahwa kami membandingkan model kami dengan YAGO2, versi yang ditingkatkan secara spasial dan temporal dari https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/. Kemudian, tiga jenis pertanyaan inti geografis diposting dan hasilnya dianalisis untuk mengevaluasi kemampuan representasi pengetahuan dari kedua model tersebut. Terakhir, evaluasi pengguna diberikan untuk memverifikasi perbandingan secara objektif.

10.7 GEOKG DAN YAGO

Struktur GeoKG dan YAGO berbeda. Meskipun Bagian sebelumnya secara singkat memperkenalkan karakteristik YAGO, perbandingan antara dua struktur yang berbeda perlu dianalisis untuk memahami perbandingan kueri berikut dan hasilnya di bagian berikutnya. Gambar 10.8 menunjukkan contoh-contoh yang disusun berdasarkan model yang berbeda.



(a) the example in YAGO structure (b) the example in GeoKG structure

Gambar 10.8. Contoh struktur model YAGO dan model GeoKG. (a) entitas, properti, dan hubungan dalam struktur YAGO; (b) elemen-elemen dalam struktur GeoKG.

Pada Gambar 10.8a, hanya ada tiga jenis elemen: entitas, properti, dan hubungan. Setiap properti tertaut ke entitas terkait melalui relasi dengan predikat. Misalnya, “Nanjing” dan “1638” memiliki hubungan bernama “startedOnDate”. Perhatikan bahwa struktur YAGO tidak memuat hubungan antar properti. Jadi, tidak ada hubungan semantik antar properti. Dengan kata lain, properti deskriptif masif dari suatu entitas tertaut ke entitas tersebut secara independen. Misalnya, dua hubungan terjadi di Nanjing dan Sungai Yangzi: “Nanjing berada di selatan Sungai Yangzi” dan “Sungai Yangzi melewati Nanjing”. Sulit untuk memahami

pengetahuan ini tanpa hubungan antar properti, sedangkan GeoKG pada Gambar 10.8b menetapkan enam elemen inti dan menghubungkan elemen-elemen ini. Dengan elemen yang lebih terintegrasi, hubungan "Nanjing berada di selatan Sungai Yangzi" dapat tergambar lebih jelas karena hubungan ini menghubungkan dua lokasi di dua negara bagian yang berbeda dari dua objek geografis. Negara bagian berbeda yang menyediakan hubungan ini terjadi pada tahun 1638 dan lokasi terkait menyediakan hubungan ini dengan deskripsi lokasi berbeda. Pengetahuan ini tidak dapat diberikan tanpa adanya hubungan antara sifat-sifat tersebut.

Konstruksi

Baik GeoKG maupun YAGO dibuat secara manual dengan menggunakan informasi tentang studi kasus evolusi pembagian administratif Nanjing. Studi kasus yang diselenggarakan oleh model YAGO adalah set rangkap tiga SPO klasik yang memiliki templat ontologi sumber terbuka. Selain itu, studi kasus yang diselenggarakan dengan model GeoKG juga disimpan oleh triple set SPO yang mengandung lebih banyak predikat. Predikat suplemen utama meliputi "isStateof", "isTimeof", "isLocationof", "isAttributeof", "isChangeof", "isRelationof", "isChangeto", dan "isRelateto". Semua predikat ini diterapkan untuk melengkapi struktur semantik model GeoKG. Dari perspektif ini, mekanisme penyimpanan yang mendasari GeoKG dan YAGO adalah sama.

Perbandingan Kemampuan Representasi Pengetahuan antara GeoKG dan YAGO

Waktu, ruang, dan atribut adalah tiga aspek yang sangat diperlukan dalam geosains. Ketiga jenis pertanyaan ini dapat didefinisikan sebagai pertanyaan standar untuk mengevaluasi apakah pengetahuan geografis yang disimpan sudah baik. Berdasarkan perbedaan antara pengetahuan faktual dan pengetahuan inferensial, setiap pertanyaan terdiri dari dua bagian. Untuk studi kasus ini, pertanyaannya ditunjukkan pada Tabel 10.2.

Tabel 10.2. Pertanyaan pada model GeoKG dan model YAGO.

Jenis Pertanyaan	**Pertanyaan Faktual**	**Pertanyaan Inferensial**
Waktu	Kapan Nanjing diberi nama ?	Kapan jiangning menjadi milik nanjing ?
Ruang	Dimana nanjingnya ?	Apa hubungan spasial antara Nanjing dan Sungai Yangzi ?
Atribut	Gaochun berada di kota manakah ?	Pembagian administrative apa yang dimiliki Nanjing ?

Pertanyaan

Pertanyaan tidak dapat ditanyakan langsung dari database GeoKG dan YAGO. Oleh karena itu, data tersebut perlu diterjemahkan ke dalam kueri SPARQL, karena GeoKG atau YAGO disimpan sebagai tripel di RDF. Misalnya, pertanyaan faktual tentang waktu dapat diterjemahkan ke dalam kueri SPARQL, seperti yang ditunjukkan pada Tabel 10.

3.

Tabel 10.3. Kueri SPARQL untuk "Kapan Nanjing diberi nama?"

Langkah	Kueri SPARQL	Arti Semantik
1	PREFIX rdf: <http://www.w3.org/2000/01/rdf-schema#>.	protokol
2	PILIH ?sWaktu DIMANA {	Konten kueri "?sTime" (waktu mulai)
3	?s rdfs:ketik :Kota.	Jenisnya adalah "Kota"
4	?s :Namakota 'Nanjing'.	Dapatkan objek geografis "Nanjing".
5	?s :memilikiNama ?o.	Dapatkan waktu ketika bernama 'Nanjing'
6	?o :startedOnDate ?sTime.	Waktu mulai
7	?o :Nama bekas ?uNama.	Kondisi kendala
8	FILTER ekspresi reguler(?uName, "^Nanjing")	Pengaturan kondisi kendala

10.8 PERBANDINGAN DAN ANALISIS

Item yang dikumpulkan YAGO dan GeoKG pada enam pertanyaan tercantum pada Tabel 10.4. Perbandingan akan dilakukan dari segi akurasi, kelengkapan, dan pengulangan.

Ketepatan

Secara umum, hasil GeoKG sedikit lebih baik dibandingkan YAGO. Kedua model tersebut dapat merespons dengan hasil yang akurat terhadap #Q1, #Q2, #Q3, #Q4, dan #Q6. Pada #Q5, hasil model YAGO menghasilkan dua item dan hasil model GeoKG menghasilkan empat item. Sebenarnya, "Zhenjiang" dan "Nanjing" dari model YAGO adalah jawaban yang menyesatkan atas pertanyaan "Di kota mana Gaochun berada?" Padahal hasil dari model GeoKG: "Zhenjiang(Gaochun, negara bagian 1949)", "Zhenjiang(Zhenjiang, negara bagian 1949)", "Nanjing(Gaochun, negara bagian 2018)" dan "Nanjing(Nanjing, negara bagian 2018) " Mirip dengan bagian depan, hasil ini berisi objek geografis dan informasi negara bagian yang relevan yang bermanfaat bagi pengguna untuk memahami hasilnya. Dari perspektif ini, informasi keadaan dari GeoKG memberikan informasi yang lebih akurat dibandingkan model YAGO.

Kelengkapan

Meskipun kedua model ini dapat memberikan hasil yang lengkap, hasil GeoKG lebih banyak mengandung integritas semantik. Di #Q6, YAGO mengembalikan 10 item: Distrik Pusat, Jiangning, Jurong, Dangtu, Luhe, Pukou, Hexian, Lishui, Gaochun, dan Qixia. Di antara divisi-divisi ini, Distrik Tengah menjadi milik Nanjing sejak tahun 1368. Jiangning, Luhe, dan Pukou menjadi milik Nanjing sejak tahun 1949. Jurong, Dangtu, dan Hexian menjadi milik Nanjing pada tahun 1949. Lishui, Gaochun, dan Qixia menjadi milik Nanjing pada tahun 2018. Seperti pertanyaannya tidak memiliki kondisi batasan waktu yang eksplisit, YAGO mengembalikan semua item, sedangkan GeoKG mengembalikan 30 item dan setiap item mencatat objek target serta objek dan keadaan geografis yang relevan. Berisi item "Distrik Tengah (Nanjing, negara bagian tahun 1368)" dan item "Distrik Pusat (Distrik Tengah, negara bagian tahun 1368)", karena terdapat hubungan yang berlawanan.

Pengulangan

Hasil GeoKG memiliki item ulangan yang lebih banyak dibandingkan hasil dari YAGO. Hasil dari YAGO terdapat item yang berulang di #Q3 dan #Q4, karena pencatatannya berulang.

Namun model GeoKG berbeda. Pada #Q2, #Q4, #Q5, dan #Q6, hasil GeoKG memiliki banyak item yang berulang; misalnya, item "1949 (Jiangning, negara bagian 1949)" dan "1949 (Nanjing, negara bagian 2018)" di #Q2. Objek target kueri "1949" juga sama. Meskipun kedua item ini bersumber dari objek geografis yang berbeda (Jiangning dan Nanjing), kedua item ini masih cukup mirip, sehingga menyebabkan lebih banyak informasi yang berlebihan bagi pengguna.

Kesimpulannya, hasil model GeoKG lebih akurat dan lengkap dibandingkan model YAGO dengan peningkatan informasi keadaan. Hal ini dapat mengurangi pengaruh pertanyaan yang tidak jelas dan memperoleh jawaban dengan makna yang lebih semantik (misalnya, objek geografis dan keadaan relevannya). Sementara itu, model GeoKG dapat menghasilkan lebih banyak hasil pasangan (misalnya, "Nanjing berada di selatan Sungai Yangzi (Nanjing, negara bagian 1368)" vs. "Nanjing berada di selatan Sungai Yangzi (Sungai Yangzi, negara bagian 1368)"), karena relasinya disimpan secara berlawanan dalam objek geografis yang berbeda.

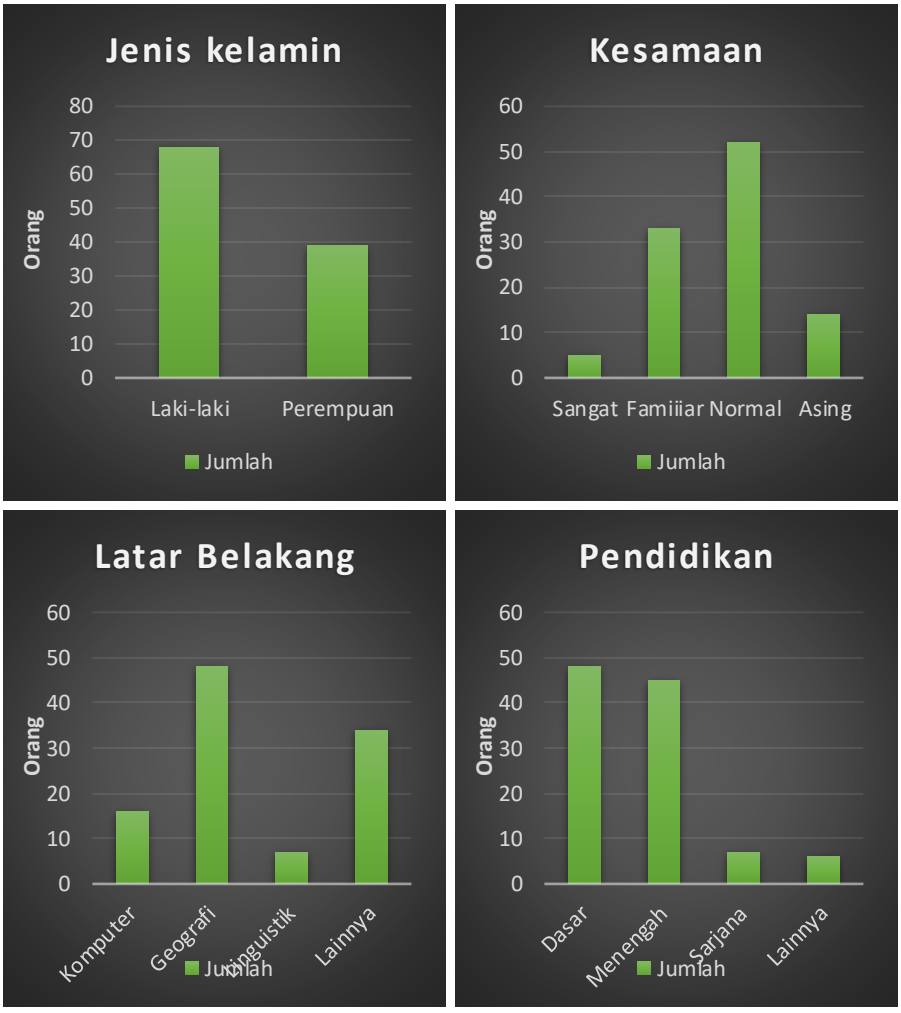
Tabel 10.4. Hasil YAGO dan GeoKG pada query SPARQL.

Jenis pertanyaan	Pertanyaan	Hasil	
		YAGO	GeoKG
Waktu	#Q1: Kapan Nanjing diberi nama?	➢ 1368	➢ 1368 (Nanjing, negara bagian tahun 1368)
	#Q2: Kapan jiangning menjadi milik nanjing ?	➢ 1949 ➢ 2018	➢ 1949 (jiangning bagian tahun 1949) ➢ 2018 (jiangning, negara bagia 2018) ➢ 1949 (Nanjing, negara bagian 2018) ➢ 2018 (Nanjing, negara bagian 2018)
Ruang	#.Q3: Dimana nanjingnya ?	➢ N32°02'38'', E188°46'43'' ➢ N32°02'38'', E188°46'43'' ➢ N32°02'38'', E188°46'43''	➢ N32°02'38'', E188°46'43'' ➢ (Nanjing, negara bagian 1368)N32°02'38'', E188°46'43'' ➢ (Nanjing, negara bagian tahun 1949) N32°02'38'', E188°46'43''
	#.Q4: Apa hubungan spasial antara Nanjing dan Sungai Yangzi ?	➢ Berada di Selatan Sungai Yangzi ➢ Sungai Yangzi melewati Nanjing ➢ Sungai Yangzi melewati Nanjing	➢ Nanjing berada di Selatan Sungai Yangzi (Nanjing negara bagian tahun 1368) ➢ Nanjing berada di Selatan Sungai Yangzi (Sungai Yangzi, negara bagian tahun 1368) ➢ Sungai Yangzi melewatinya Nanjing

			(Nanjing, negara bagian tahun 1949) ➢ Sungai Yangzi melewati Nanjing (Sungai Yangzi, negara bagian tahun 1949) ➢ Sungai yangzi melewatinya Nanjing(Nanjing, negara bagian 2018) ➢ Sungai Yangzi melewati nannjing (Sungai Yangzi negara bagian 2018)
Atribut	#Q5: Gaochun berada di kota mana ?	➢ Zhenjiang ➢ Nanjing	➢ Zhenjiang (gaochun, negara bagian tahun 1949) ➢ Zhenjiang (Zhenjiang negara bagian tahun 1949) ➢ Nanjing (gaochun, negara bagian tahun 2018) ➢ Nanjing (Nanjing, negara bagian 2018)
	#Q6: apa pembagian administrative milik Nanjing ?	➢ Distrik Tengah ➢ Jiangning ➢ Jurong ➢ Dangtu ➢ Luhe ➢ Pukou ➢ Hexian ➢ Lishui ➢ Gaochun ➢ Qixia	➢ Distrik pusat (Nanjing, negara bagian 1368) ➢ Distrik pusat(Nanjing, negara bagian tahun 1949) ➢ Distrik pusat (Nanjing, negara bagian tahun 2018) ➢ Jiangning (Nanjing,negara bagian tahun1949) ➢ Dangtu (Nanjing, negara bagian tahun 1949) ➢ Luhe (Nanjing, negara bagian 1949) ➢ Luhe (Nanjing, negara bagian 2018) ➢ Pokou (Nanjing, negara bagian tahun 1949) ➢ Pukou (Nanjing negara bagian 2018) ➢ Hexian(Nanjing, negara bagian tahun 1949) ➢ Lishui(Nanjing negara bagian 2018) ➢ Gaochun (Nanjing, negara bagian 2018) ➢ Qixia (Nanjing, negara bagian 2018) ➢ Lebih banyak item

Evaluasi Pengguna

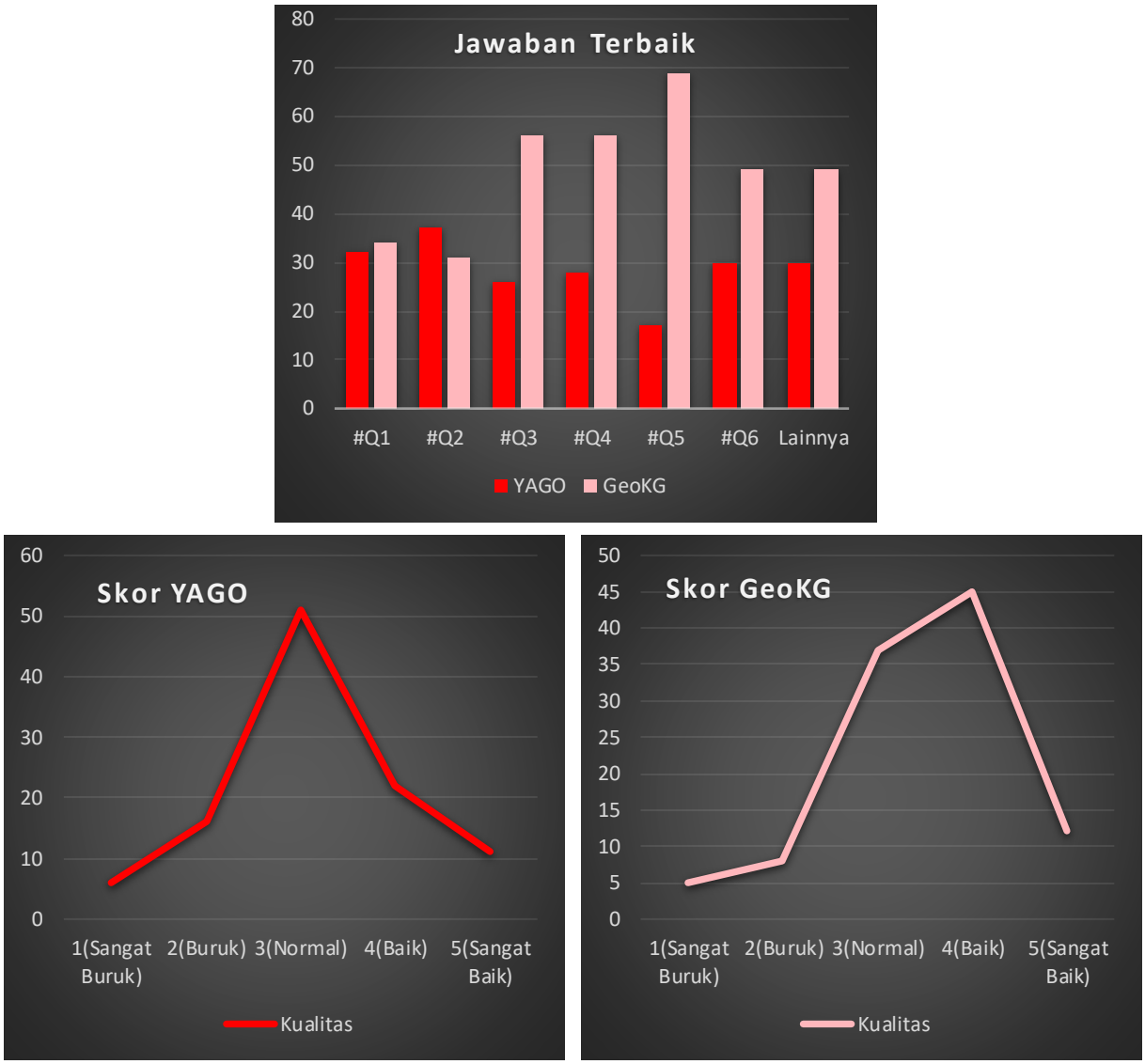
Survei kuesioner online juga diberikan untuk memverifikasi hasil analisis komparatif. Kuesioner dibagi menjadi delapan bagian. Bagian pertama adalah survei informasi dasar yang menanyakan empat aspek informasi kepada individu (gender, keakraban dengan wilayah penelitian, latar belakang, dan tingkat pendidikan). Statistik informasi dasar ini ditunjukkan pada Gambar 10.9. pada bagian sebelumnya berhubungan dengan pertanyaan #Q1–#Q6 dan menanyakan pertanyaan tentang jawaban terbaik, keakuratan, kelengkapan, dan pengulangan. Ringkasan pertanyaan termasuk evaluasi keseluruhan, skor pada YAGO dan skor pada GeoKG pada berbagai aspek. Skor ditetapkan sebesar 1–5 sesuai dengan sangat buruk, buruk, normal, baik, dan sangat baik, dan setiap kelompok skor mencakup skor keseluruhan, skor akurasi, skor kelengkapan, dan skor pengulangan. Ada 106 masukan valid yang akhirnya kami terima.



Gambar 10.9. Statistik dari empat jenis informasi dasar survei.

Gambar 10.10 menunjukkan jawaban terbaik pada #Q1–#Q6 dan skor keseluruhan YAGO dan GeoKG. Pada histogram jawaban terbaik, hasil keseluruhan menunjukkan 54,72% individu mendukung GeoKG, lebih tinggi 23,59% dibandingkan YAGO sebesar 31,13%. Secara khusus, kuantitas #Q1 dan #Q2 cukup dekat namun kuantitas #Q3–#Q6 tidak. Kuantitas

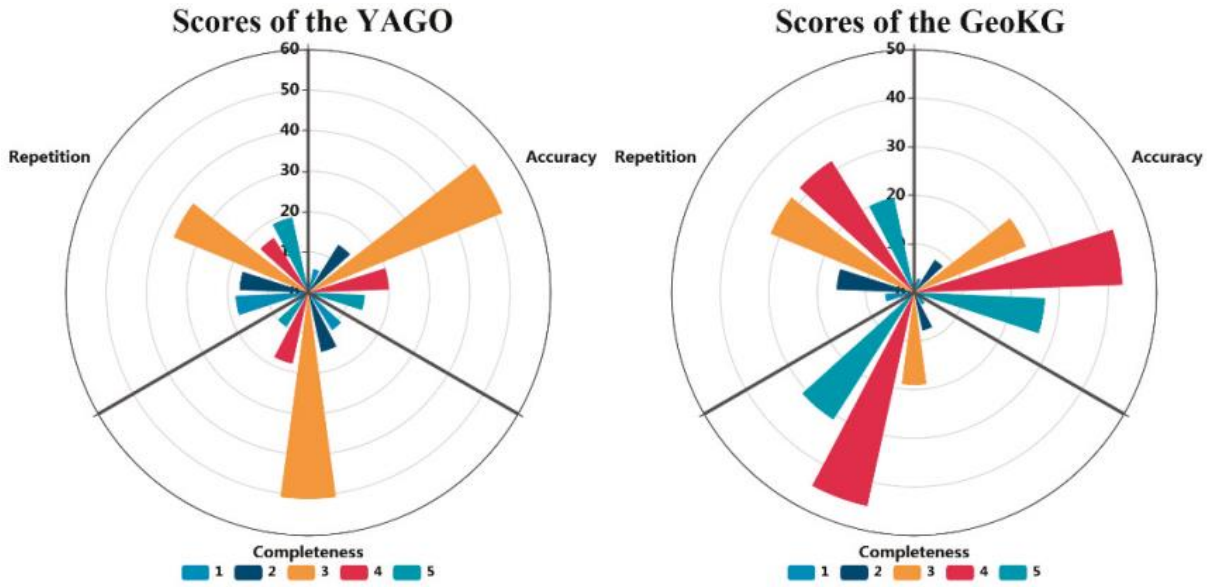
GeoKG jauh lebih tinggi dibandingkan YAGO di antara empat pertanyaan terakhir, terutama di #Q5. Grafik garis skor keseluruhan pada YAGO dan GeoKG juga menunjukkan bahwa evaluasi GeoKG lebih baik dibandingkan YAGO. Didapatkan peningkatan skor rata-rata sebesar 7,8% dari YAGO (3,15) ke GeoKG (3,49).



Gambar 10.10. Jawaban terbaik pada #Q1–#Q6 dan skor keseluruhan YAGO dan GeoKG.

Dari sudut pandang subaspek (akurasi, kelengkapan, dan pengulangan), besaran yang berbeda dapat langsung menunjukkan skor dari YAGO dan GeoKG. Besaran yang berbeda-beda menunjukkan kemampuan dari model (detailnya pada Gambar 10.11). Hampir ketiga aspek YAGO memperoleh skor 3, sedangkan GeoKG berbeda: skor 4 untuk akurasi, skor 4–5 untuk kelengkapan, dan skor 3–4 untuk pengulangan. Membandingkan skor-skor ini, terlihat bahwa terdapat sedikit peningkatan pada akurasi dari skor rata-rata 3,11 di YAGO menjadi skor rata-rata 3,78 di GeoKG. Peningkatan luar biasa terlihat pada kelengkapan jawaban dari skor rata-rata 2,99 di YAGO menjadi skor rata-rata 3,87 di GeoKG. Selain itu, GeoKG juga

memperoleh repetisi yang lebih tinggi dari skor rata-rata 3,01 di YAGO menjadi skor rata-rata 3,42 di GeoKG.



Gambar 10.11. Peta mawar dari sejumlah aspek berbeda pada YAGO dan GeoKG.

Singkatnya, jawaban dari GeoKG membuat perbaikan terhadap jawaban YAGO. Evaluasi pengguna secara objektif memverifikasi analisis dan secara khusus menunjukkan jawaban yang jelas. Terlihat bahwa perbaikan utama GeoKG ada pada #Q3–#Q6 yaitu pertanyaan spasial dan atribut. Jawaban atas pertanyaan-pertanyaan ini memerlukan lebih banyak informasi keadaan terkait dan informasi temporal, yang memerlukan hubungan antar elemen (Gambar 10.8). Inilah alasan mengapa GeoKG lebih baik dari YAGO. Selain itu, GeoKG berisi lebih banyak informasi redundansi daripada YAGO karena elemen relasinya bersifat dua arah. Hal ini dapat menjadi fokus penelitian lebih lanjut yang berkelanjutan mengenai indeks dan penerapannya di masa depan.

10.9 RINGKASAN

Mengingat banyaknya perhatian yang diberikan pada representasi pengetahuan geografis, pada bab ini difokuskan pada pengembangan representasi pengetahuan geografis saat ini. Kami menganalisis masalah representasi pengetahuan geografis saat ini dan menemukan bahwa dua masalah harus diperbaiki: elemen representasi pengetahuan geografis dan penambahan operator konstruksi DL.

Mengikuti ide dasar dari enam pertanyaan inti geografis, kami merancang model konseptual yang disebut GeoKG berdasarkan enam elemen di sekitar pertanyaan geografis, kemudian melengkapi operator konstruksi DL dan akhirnya memberikan formalisasi model dengan operator tersebut. Selain itu, kasus evolusi pembagian administratif Nanjing diformalkan dan diilustrasikan. Kemudian, grafik pengetahuan dibangun dengan model GeoKG dan model YAGO dengan menggunakan studi kasus. Setelah menetapkan sekelompok pertanyaan geografis standar, hasil kueri akhirnya dibandingkan. Hasil penelitian

menunjukkan bahwa hasil GeoKG lebih akurat dan lengkap dibandingkan hasil YAGO yang diverifikasi melalui evaluasi pengguna berikut ini. Perbandingan ini menunjukkan model GeoKG menampilkan kemampuannya untuk mengatur pengetahuan geografis di komputer dan merupakan model yang menjanjikan dan kuat untuk representasi pengetahuan geografis.

BAB 11
INFRASTRUKTUR SIBER DALAM DATA BESAR IKLIM DI THREDDS

Memahami perilaku iklim di masa lalu, masa kini, dan perubahan memerlukan kolaborasi erat dari sejumlah besar peneliti dari berbagai bidang ilmiah. Saat ini, kolaborasi interdisipliner yang diperlukan sangat dibatasi oleh kesulitan dalam menemukan, berbagi, dan mengintegrasikan data iklim karena ukuran data yang semakin meningkat. Pada bab ini membahas metode dan teknik untuk memecahkan masalah yang saling terkait yang dihadapi saat mentransmisikan, memproses, dan menyajikan metadata untuk data Observasi dan Pemodelan Sistem Bumi (ESOM) yang heterogen. Solusi berbasis infrastruktur siber diusulkan untuk memungkinkan pembuatan katalog yang efektif dan pencarian dua langkah pada kumpulan data iklim besar dengan memanfaatkan teknologi layanan web tercanggih dan merayapi pusat data yang ada. Untuk memvalidasi kelayakannya, kumpulan data besar yang disajikan oleh UCAR THREDDS Data Server (TDS), yang menyediakan data ESOM tingkat Petabyte dan memperbarui ratusan terabyte data setiap hari, digunakan sebagai kumpulan data studi kasus. Alur kerja lengkap dirancang untuk menganalisis struktur metadata di TDS dan membuat indeks untuk parameter data. Model registrasi yang disederhanakan yang mendefinisikan informasi konstan, membatasi informasi sekunder, dan memanfaatkan koherensi spasial dan temporal dalam metadata telah dibangun. Model ini memperoleh strategi pengambilan sampel untuk bot perayap web serentak berkinerja tinggi yang digunakan untuk mencerminkan metadata penting dari arsip data besar tanpa membebani sumber daya jaringan dan komputasi. Model metadata, crawler, dan layanan katalog yang sesuai standar membentuk infrastruktur siber pencarian tambahan, yang memungkinkan para ilmuwan untuk mencari kumpulan data iklim besar hampir secara real-time. Pendekatan yang diusulkan telah diuji pada UCAR TDS dan hasilnya membuktikan bahwa pendekatan ini mencapai tujuan desainnya dengan setidaknya meningkatkan kecepatan perayapan sebanyak 10 kali lipat dan mengurangi metadata yang berlebihan dari 1,85 gigabyte menjadi 2,2 megabyte, yang merupakan terobosan signifikan untuk membuat terobosan saat ini. sebagian besar server data iklim yang tidak dapat dicari dapat dicari.

11.1. PENDAHULUAN

Infrastruktur siber memainkan peran penting dalam kegiatan penelitian iklim saat ini. Para ilmuwan iklim mencari, menelusuri, memvisualisasikan, dan mengambil data spasial menggunakan sistem web setiap hari, terutama karena volume data dari observasi dan simulasi model tumbuh dalam jumlah besar sehingga perangkat pribadi tidak dapat menampung seluruhnya. Tantangan big data dalam hal volume, kecepatan, variasi, kebenaran, dan nilai (5V), telah mendorong penelitian geosains menjadi upaya yang lebih kolaboratif yang melibatkan banyak penyedia data observasi, pengembang infrastruktur siber, pemodel, dan pemangku kepentingan informasi. Ilmu pengetahuan iklim telah berkembang selama beberapa dekade dan menghasilkan produk data berukuran puluhan petabyte,

termasuk observasi stasioner, hindcast, dan analisis ulang, yang disimpan di pusat data terdistribusi di berbagai negara di seluruh dunia. Individu atau kelompok kecil ilmuwan menghadapi tantangan besar ketika mereka berupaya menemukan data yang mereka perlukan secara efisien. Saat ini, sebagian besar ilmuwan memperoleh pengetahuan tentang kumpulan data melalui konferensi, rekomendasi rekan kerja, buku teks, dan mesin pencari. Mereka menjadi sangat akrab dengan kumpulan data yang mereka gunakan, dan setiap kali mereka ingin mengambil data, mereka langsung membuka situs web kumpulan data untuk mengunduh data yang sesuai dengan rentang waktu dan spasial yang diminta. Namun, rutinitas ini kurang berkelanjutan karena sensor/kumpulan data menjadi lebih bervariasi, model lebih sering berevolusi, dan data baru yang berkaitan dengan penelitian tersedia di tempat lain.

Dalam sebagian besar skenario, metadata adalah informasi pertama yang dilihat peneliti, sebelum mereka mengakses dan menggunakan data pengamatan dan pemodelan Bumi sebenarnya yang dijelaskan oleh metadata. Berdasarkan metadata, mereka memutuskan apakah data aktual akan berguna dalam penelitian mereka atau tidak. Untuk data spasial besar, metadata adalah komponen kunci yang mendukung semua jenis operasi sehari-hari pengguna, seperti pencarian, pemfilteran, penjelajahan, pengunduhan, tampilan, dll. Saat ini, dua masalah mendasar dalam mengakses dan menggunakan data spasial besar adalah volume metadata dan kecepatan pemrosesan metadata. Melalui penyelidikan manual terhadap repositori data Unidata THREDDS (sumber metadata yang kami ambil sebagai contoh pola penyimpanan geodata yang khas), terungkap bahwa sebagian besar metadata sangat berlebihan. Sebagian besar catatan metadata berisi informasi yang identik dan hanya bidang utama yang mewakili karakteristik spasial dan temporal yang diperbarui secara berkala. Namun, terdapat pola yang teratur mengenai bagaimana informasi redundan disusun dan bagaimana informasi baru ditambahkan ke repositori—namun, pola tersebut bervariasi menurut hierarki organisasi data dan berubah sesuai dengan jenis data yang dikirimkan (misalnya: Stasiun radar vs. (pengamatan satelit vs. keluaran model prakiraan reguler).

Untuk mengatasi tantangan pencarian data besar ini, kita harus menghadapi permasalahan praktis di bidang ini model informasi, kualitas informasi, dan teknis implementasi sistem informasi. Studi kami mengikuti hubungan antara tantangan ilmiah mendasar dan implementasi sistem informasi geosains yang ada. Studi ini bertujuan untuk membangun model katalogisasi yang mampu mendeskripsikan metadata heterogen secara real-time sekaligus mengurangi volume data dan memungkinkan pencarian dalam repositori data Big Earth. Model ini dapat digunakan untuk merepresentasikan data redundan secara efisien dalam repositori metadata asli dan untuk melakukan kompresi informasi lossless untuk penyimpanan dan pencarian yang ringan dan efisien. Model ini dapat memperkecil metadata dalam jumlah besar (tanpa mengorbankan kompleksitas informasi atau variasi yang tersedia dalam repositori asli) dan mengurangi beban komputasi dalam pencarian metadata. Model ini mendefinisikan dua jenis objek: Koleksi dan butiran. Ini juga menentukan siklus hidup dan hubungannya dengan data repositori THREDDS hulu. Koleksi berisi metadata konten (judul,

deskripsi, kepengarangan, informasi variabel/band, dll.). Setiap koleksi berisi satu atau lebih butiran. Setiap butiran hanya berisi metadata tingkat spatiotemporal.

Hal baru dari penelitian ini adalah mengubah repositori pusat data lama menjadi layanan katalog ringan yang fleksibel, yang lebih mudah dikelola dengan menyediakan kemampuan pencarian kumpulan data berukuran petabyte. Pekerjaan ini memberikan referensi penting bagi orang-orang yang mengoperasikan pusat data iklim besar dan memberikan saran mengenai perbaikan lebih lanjut pada pusat data iklim operasional tersebut agar dapat melayani komunitas ilmu iklim dengan lebih baik

Studi yang dijelaskan dalam bab ini merupakan upaya untuk berkontribusi pada upaya ilmiah global dalam memahami dan memprediksi dampak perubahan iklim. Memahami perubahan iklim dan dampaknya memerlukan pemahaman Bumi sebagai sistem yang kompleks dengan perilaku yang muncul dari interaksi dan umpan balik yang terjadi pada berbagai skala temporal dan spasial. Namun, kemajuan baru dalam studi ini terhambat oleh tantangan kolaborasi interdisipliner dan kesulitan kolaborasi data dan informasi. Kesulitan dalam kolaborasi informasi dapat dipahami dalam kaitannya dengan permasalahan big data yang sudah berlangsung lama dalam hal keragaman (kompleksitas) dan volume/kecepatan.

11.2. METADATA DAN KOLABORASI DATA

Metadata adalah alat yang ampuh untuk menghadapi tantangan big data. Kami mendiskusikan latar belakang metadata dan interoperabilitas katalog metadata sebagai komponen penting dari infrastruktur siber canggih yang kami impikan.

Topik metadata telah didekati oleh dua tradisi ilmiah yang berbeda. Memahaminya membantu kami memperjelas pendekatan kami terhadap metadata dalam infrastruktur siber. Ilmuwan informasi perpustakaan telah menjelaskan pendekatan kontrol bibliografi metadata. Prinsip bibliografi memungkinkan pengguna informasi untuk mendeskripsikan, menemukan, dan mengambil entitas yang mengandung informasi. Unit metadata dasar adalah "pengganti informasi" yang kegunaannya dapat ditemukan (menurut penulis, judul, dan subjek), secara akurat mendeskripsikan objek informasi (data metadata) dan mengidentifikasi cara menemukan objek tersebut. Pandangan metadata yang kedua (pelengkap) berasal dari disiplin ilmu komputer dan disebut pendekatan manajemen data. Data yang kompleks dan heterogen (tekstual, grafis, relasional, dll.) tidak dipisahkan menjadi unit informasi, namun dijelaskan oleh model dan arsitektur data yang mewakili "informasi tambahan yang diperlukan agar data berguna". Perbedaan utamanya adalah pendekatan bibliografi bekerja dengan entitas informasi berbeda dengan tipe terbatas, sedangkan pendekatan manajemen data bekerja dengan model struktur data/informasi dan hubungannya.

Perbedaan antara pendekatan bibliografi dan manajemen data ini penting dalam konteks upaya standardisasi metadata yang sedang berlangsung. Pendekatan kedua tidak kondusif untuk standardisasi karena model pengelolaan data sama kompleks dan heterogennya dengan struktur data yang dimodelkan. Oleh karena itu, sesuai dengan standar yang ada, metadata yang tersedia saat ini untuk kumpulan data iklim berukuran besar mengikuti pendekatan pertama, yang memberikan informasi bibliografi dan tidak

mendeskripsikan struktur data dengan cara yang memungkinkan kapasitas baru infrastruktur siber yang canggih. Bab kami menjelaskan upaya untuk melengkapi dan mengubah metadata bibliografi yang ada dengan model manajemen metadata khusus yang menghasilkan aplikasi baru untuk data yang sudah ada. Standarisasi metadata merupakan prasyarat interoperabilitas yang merupakan prasyarat untuk membangun sistem informasi terdistribusi yang mampu menangani data sistem bumi yang kompleks.

Interoperabilitas, Katalog Data, Sistem Geoinformasi, dan THREDDS

Kolaborasi data dan informasi lintas disiplin ilmu sangat penting bagi ilmu kebumian tingkat lanjut. Sayangnya, tidak ada praktik terpadu yang kuat untuk perekaman, penyimpanan, transmisi, dan pemrosesan data yang diikuti oleh seluruh komunitas ilmiah. Bidang dan tradisi yang berbeda mempunyai format data, perangkat lunak, dan prosedur manajemen data pilihannya masing-masing. Namun, studi sistem bumi umumnya bekerja dengan data yang mengikuti format geografis-temporal. Semua data dapat disimpan secara bermakna pada kisi dimensi 4D (3 spasial dan satu temporal). Kesamaan dasar ini telah mengilhami upaya standardisasi dengan tujuan memungkinkan interoperabilitas dan kolaborasi yang lebih luas.

Mengikuti perkembangan organik dari masyarakat, upaya standardisasi kini dipimpin oleh Open Geospatial Consortium (OGC) dan Organisasi Internasional untuk Komite Teknis Standardisasi 211 (ISO TC 211) dan telah menghasilkan standar yang berhasil di dua bidang yang relevan bagi kami. Pertama adalah definisi NetCDF sebagai salah satu format data standar untuk menyimpan data geografis. Yang kedua adalah standardisasi metadata. Upaya-upaya tersebut sangat relevan dengan penelitian kami dan dibahas lebih lanjut di bagian Pekerjaan Terkait. Sebagai latar belakang, penting untuk disebutkan bahwa model metadata geografis standar yang dikembangkan oleh OGC masih terus mengembangkan kemampuan untuk mendeskripsikan data besar yang heterogen, bervolume tinggi, atau berkecepatan tinggi yang sedang kita pelajari. Standar metadata seri OGC/ISO 19* yang umum digunakan memiliki fitur relasional yang relatif terbatas (hanya agregasi) dan, dalam repositori yang kami pelajari, setiap rekaman metadata yang dikodekan XML berisi sebagian besar informasi berlebihan (misalnya, dua objek metadata yang mewakili dua gambar dari sebuah sensor tunggal sebagian besar berisi informasi duplikat yang menggambarkan karakteristik sensor).

Namun, ada beberapa bidang pekerjaan yang dilakukan ISO TC 211 untuk mengatasi masalah ini dan menyarankan tren untuk memperluas penerapan model metadata standar dan integrasi lebih banyak variasi informasi. Model metadata geografis standar dikembangkan bersama dengan model registrasi katalog terdistribusi standar yang berjudul Layanan Katalog untuk Web (CSW). Standar CSW dikenal luas dan banyak penyedia data sistem Bumi menawarkan beberapa informasi tentang kepemilikan data mereka melalui antarmuka CSW. Namun, standar CSW juga kurang cocok untuk mendukung studi kolaboratif big data untuk sistem Bumi. CSW mengikuti model metadata OGC dasar dengan cara yang menyulitkan untuk menangkap struktur berharga dan semantik dari penyimpanan data yang ada tanpa menyimpan informasi yang sangat berlebihan—yang menghabiskan sumber daya komputasi tanpa memanfaatkan nilai sebenarnya dari data besar Bumi yang besar dan kompleks. Namun,

metadata OGC yang disimpan di CSW adalah standar yang ada yang tidak hanya mengatur praktik distribusi data, tetapi juga cara berpikir peneliti tentang kolaborasi data.

Item berikutnya yang digunakan dalam penelitian ini adalah UCAR Unidata THREDDS Data Server (TDS). University Corporation for Atmospheric Research (UCAR) Unidata adalah komunitas kolaborasi data geosains dari beragam lembaga penelitian dan pendidikan. Ini memberikan data sistem Bumi heterogen secara real-time yang menjadi target penelitian ini. THREDDS adalah Layanan Data Terdistribusi Lingkungan Real-Time Tematik Unidata. TDS adalah server web yang menyediakan metadata dan akses data untuk kumpulan data ilmiah bagi para peneliti iklim. TDS menyediakan layanan katalog hierarki dasar yang tidak dapat dicari dan tidak mendukung standar CSW. Namun, hal ini mendukung standar metadata geografis OGC—walaupun tidak secara konsisten dan komprehensif. Agar data yang dihosting oleh TDS dapat dicari, metadata TDS harus disalin ke server lain dan katalog yang dapat dicari harus dibuat untuk metadata tersebut. Tugas ini dilakukan oleh perayap web khusus yang dikembangkan oleh penelitian ini. Studi ini mencoba untuk membangun infrastruktur yang ada dengan sumber daya yang tersedia dan keterbatasannya untuk memberikan kemampuan baru. Keterbatasan sistem yang ada ada dua. Pertama adalah keterbatasan model pendaftaran metadata CSW (model ini tidak secara alami mendukung pendaftaran informasi tentang siklus hidup metadata atau informasi agregasi yang cukup rinci), dan kedua adalah ketidaklengkapan informasi dalam metadata yang disediakan oleh THREDDS. Studi ini mencoba untuk menghapus batasan tersebut dengan terlebih dahulu melakukan interpolasi informasi untuk meningkatkan kualitas model metadata yang ada dan kemudian dengan memperluas model tersebut untuk memberikan kemampuan tingkat lanjut. Panduan ini mendemonstrasikan cara mengintegrasikan metadata TDS dengan perangkat lunak CSW dan mengusulkan beberapa solusi praktis yang mengatasi keterbatasan model registrasi metadata CSW. Kami melakukan ini untuk menunjukkan bahwa peningkatan kemampuan metadata dan katalog juga dapat mengurangi tantangan big data dalam variabilitas, volume,
dan kecepatan.

11.3 PERMODELAN METADATA

Pada bagian ini menyatukan beberapa bidang kerja yang ada untuk menghadapi permasalahan dalam pengintegrasian dan pencarian kumpulan data ilmu iklim yang luas dan beragam. Penelitian yang ada di bidang pemodelan metadata, interoperabilitas informasi geografis, katalog geografis, perayapan informasi web, dan pengindeksan pencarian memberikan landasan bagi upaya kami untuk mendemonstrasikan dan mengevaluasi kemampuan infrastruktur siber data iklim tingkat lanjut.

Ada banyak penelitian yang mengeksplorasi hubungan mendasar antara model metadata dan kemampuan informasi. Terdapat beragam penelitian di bidang lain yang menangani permasalahan dasar yang sama dan menunjukkan bahwa pembuatan model metadata baru dapat digunakan sebagai metode untuk memecahkan tantangan informasi. Misalnya, Spéry dkk. telah mengembangkan model metadata untuk menggambarkan garis keturunan perubahan objek geografis dari waktu ke waktu. Mereka menggunakan grafik

asiklik langsung dan serangkaian operasi dasar untuk membangun model mereka. Model ini mendukung aplikasi baru dalam menanyakan data kadaster historis dan meminimalkan ukuran informasi metadata geografis. Pemodelan metadata spatiotemporal dapat digeneralisasikan sebagai deskripsi objek dalam ruang dan waktu, dan hubungan antar objek dipahami sebagai aliran informasi, energi, dan material untuk memodelkan evolusi objek yang saling bergantung dalam suatu sistem. Pemodelan asal ("riwayat derivasi produk data mulai dari sumber aslinya") adalah bagian penting dari studi metadata. Model metadata dan sistem informasi yang ada telah diperluas secara eksperimental dengan kemampuan pemodelan asal untuk memungkinkan visualisasi riwayat data dan analisis alur kerja yang memperoleh produk data yang digunakan oleh para ilmuwan. Eksperimen untuk mengkonsep ulang metadata sebagai praktik "manajemen pengetahuan" menghasilkan model metadata yang dapat mendukung kebutuhan pengambilan keputusan spasial dengan mengidentifikasi masalah hubungan entitas, integritas, dan presentasi. Model metadata yang diusulkan memungkinkan penyampaian informasi yang lebih kompleks tentang data spasial. Model metadata ini memungkinkan pembuatan aplikasi informasi geografis asli, bernama Florida Marine Resource Identification System, yang memperluas penggunaan data lingkungan dan sipil yang ada untuk memberdayakan pengguna dengan pengetahuan tingkat tinggi untuk analisis dan perencanaan. Melihat ke luar domain geografis, kami masih mengamati bahwa pengenalan pendekatan dan model metadata khusus memungkinkan pengembangan kemampuan baru.

Standardisasi Metadata Geografis, Interoperabilitas, dan Katalogisasi

Keberagaman model dan format metadata yang dikembangkan oleh penelitian telah memungkinkan sistem geoinformasi baru yang kuat, namun juga menimbulkan serangkaian masalah baru dalam penggunaan kembali data dan interoperabilitas. Organisasi penelitian, administrasi, dan bisnis publik dan swasta telah mengumpulkan simpanan geoinformasi dan data yang terus bertambah, namun data ini belum menjadi lebih mudah untuk ditemukan dan diakses oleh pengguna di luar yurisdiksi organisasi terbatas. Hal ini menyebabkan pemborosan sumber daya secara signifikan dan duplikasi upaya bagi produsen dan konsumen data. Katalogisasi menjadi semakin menantang karena heterogenitas ini. Sebagai responnya, infrastruktur data spasial baru telah dikembangkan. Mereka telah berupaya untuk mengintegrasikan dan menstandardisasi beberapa model metadata dan mengembangkan model kosa kata semantik bersama untuk memungkinkan penemuan dengan menggunakan model metadata "perpustakaan digital". Dalam proses ini, tantangan interoperabilitas sintaksis dan semantik telah diidentifikasi. Pengoperasian sintaksis mengacu pada portabilitas informasi—kemampuan sistem untuk bertukar informasi. Interoperabilitas semantik mengacu pada pengetahuan domain yang memungkinkan layanan informasi memahami cara menggunakan data dari sistem lain secara bermakna.

Berbagai teknik untuk mencapai interoperabilitas metadata telah dieksplorasi. Dua kelompok teknik yang terkait dapat diidentifikasi. Pendekatan yang satu berupaya menciptakan model standar dan universal, sedangkan pendekatan yang lain menciptakan pemetaan antara beberapa representasi metadata dari data yang sama. Transformasi antara beberapa model metadata mengharuskan heterogenitas sintaksis, struktural, dan semantik

dapat diselaraskan. Rekonsiliasi dilakukan dengan teknik yang disebut penyeberangan metadata. Penyeberangan adalah "pemetaan elemen, semantik, dan sintaksis dari satu skema metadata ke skema metadata lainnya". Setelah pemetaan dikembangkan, pemetaan tersebut dapat digunakan untuk menerapkan beberapa skema metadata ke data yang ada.

Kemungkinan interoperabilitas telah ditingkatkan melalui upaya yang dipimpin oleh Organisasi Internasional untuk Komite Teknis Standardisasi 211 (ISO TC 211) untuk menstandardisasi representasi metadata. Ini memperkenalkan serangkaian standar metadata geografis ISO 19* untuk mendeskripsikan informasi geografis melalui metadata. Standar ini mendefinisikan elemen metadata wajib dan opsional serta hubungan antar elemen. Misalnya, cakupan spatiotemporal, kepenulisan, dan deskripsi umum kumpulan data diwajibkan dan direkomendasikan oleh standar. Jenis informasi lain seperti pengurutan kumpulan data dalam koleksi, agregasi, dan data relasional lainnya bersifat opsional dalam standar. Seri standar ISO 19* juga menyediakan skema XML untuk representasi metadata dalam XML.

Melihat kerja interoperabilitas metadata yang ada, kami melihat terulangnya masalah serupa seperti keragaman representasi metadata dan kompleksitas pemetaan di antara keduanya. Beberapa penulis membahas tantangan praktis dalam mengembangkan perangkat lunak dan sistem penerjemahan. Terdapat banyak upaya dan hasil studi yang memajukan tujuan interoperabilitas dengan mengidentifikasi pemahaman utama tentang tantangan interoperabilitas dan menunjukkan sistem, layanan, dan model yang mengatasi tantangan bersama. Pekerjaan kami berupaya untuk mempertahankan kemajuan interoperabilitas yang ada sambil menjajaki kemungkinan memperluas model metadata yang ada untuk mendukung kemungkinan baru penggunaan data yang ada.

Metadata standar sering kali disimpan dan tersedia menggunakan layanan katalog. Katalog memungkinkan pengguna untuk menemukan metadata menggunakan kueri yang menggambarkan karakteristik informasi spasial, temporal, tekstual, dan lainnya yang diinginkan dari data yang dicari. Layanan Katalog OGC untuk Web (CSW) adalah salah satu model katalog yang banyak digunakan dalam domain geosains untuk menggambarkan kepemilikan informasi geografis.

Pemanenan dan Perayapan Web

Salah satu kapasitas penting infrastruktur cyber metadata adalah kemampuan untuk mengintegrasikan metadata dari repositori web jarak jauh. Proses menemukan dan mengimpor data tertaut web ke dalam repositori metadata disebut "perayapan" dan dilakukan dengan menggunakan sistem perangkat lunak yang disebut "perayap web metadata". Perayap web adalah program komputer yang menjelajahi web dengan "cara yang metodis, otomatis, atau teratur". Crawler adalah bot internet, ini adalah program yang secara mandiri dan sistematis mengambil data dari world wide web. Secara otomatis menemukan dan mengumpulkan berbagai sumber daya secara teratur dari internet sesuai dengan seperangkat aturan yang ada di dalamnya. Patil dan Patil merangkum arsitektur umum web crawler dan juga memberikan definisi beberapa jenis web crawler. Perayap terfokus adalah jenis yang dirancang untuk menghilangkan pengunduhan data web yang tidak perlu dengan memasukkan algoritme untuk memilih tautan mana yang harus diikuti. Perayap tambahan

terlebih dahulu memeriksa perubahan dan pembaruan pada halaman sebelum mengunduh data lengkapnya. Ini tentu melibatkan tabel indeks tanggal dan waktu pembaruan halaman. Kami mengikuti dua strategi ini dalam desain crawler kami. Penulis juga menguraikan strategi umum untuk mengembangkan crawler terdistribusi dan paralel. Crawler kami berjalan pada satu mesin, namun kami menggunakan model proses multithread dengan mekanisme antrian bersama—strategi paralelisasi umum yang diidentifikasi oleh penulis.

Sebuah tinjauan yang cukup baru dikumpulkan oleh Desai et al. menunjukkan bahwa penelitian perayap web adalah bidang pekerjaan yang aktif—namun, sebagian besar pekerjaan ini difokuskan pada kebutuhan konstruksi indeks mesin pencari web secara umum. Terdapat area penelitian yang disebut "perayapan vertikal" yang membahas masalah perayapan data web non-tradisional: item berita, daftar belanja online, gambar, audio, video. Tampaknya tidak ada publikasi apa pun mengenai perayapan metadata sistem Bumi yang heterogen secara efisien.

Terdapat penelitian substansial sebelumnya yang menunjukkan kelayakan perayapan metadata ini. Sebuah bab baru-baru ini merangkum keadaan terkini. Li dkk. menyajikan sistem perayapan dan pencarian metadata sistem Bumi yang heterogen bernama PolarHub—alat perayapan web yang mampu melakukan pencarian skala besar dan perayapan data geografis terdistribusi. Ia menggunakan mesin pencari web tekstual (Google) yang ada untuk menemukan layanan data geografis yang sesuai standar OGC. Ini menyajikan antarmuka interaktif yang memungkinkan pengguna menemukan variasi dan keragaman katalog dan layanan data terkait. Ia memiliki arsitektur sistem perangkat lunak multi-thread terdistribusi yang canggih. PolarHub menunjukkan bahwa data dari banyak sumber dapat disajikan di satu tempat. Namun, ini tidak menyajikan kumpulan data, hanya titik akhir yang harus dijelajahi sendiri oleh pengguna. Itu tidak mengunduh, meringkas, atau menyelaraskan metadata yang disimpan di katalog jarak jauh. Hal ini menunjukkan kelayakan infrastruktur siber yang mengintegrasikan berbagai data berdasarkan standar yang dapat dioperasikan namun tidak membahas tantangan volume dan kecepatan data yang muncul ketika perayapan yang lebih dalam dan lebih lengkap dilakukan. Pengguna PolarHub dapat menemukan sejumlah besar katalog dan layanan yang berisi, misalnya, data "suhu air permukaan" tetapi mereka tidak dapat menggunakan perayap metadata yang mengikuti strategi hub katalog ini untuk menemukan kumpulan data yang menyimpan "suhu air permukaan dalam rentang spasial dan temporal X dengan Y resolusi spasial dan temporal".

Strategi pelengkap dibahas oleh Pallickara dkk., yang menyajikan sistem perayapan metadata bernama GLEAN, yang menyediakan katalog web baru untuk data atmosfer berdasarkan ekstraksi metadata terperinci dari pengumpulan data atmosfer skala besar yang ada. Ini memecahkan masalah volume data dengan memperkenalkan skema metadata baru berdasarkan kumpulan data sintetis khusus yang mewakili kumpulan (atau subkumpulan atau perpotongan) dari beberapa kumpulan data yang ada. Hal ini sangat mengurangi overhead metadata dan memungkinkan penemuan dan akses dataset tertentu yang berkinerja tinggi dan tepat di dalam penyimpanan data atmosfer yang luas. Tidak seperti PolarHub, GLEAN menghindari tantangan variasi data dengan membatasi pemrosesannya pada satu jenis

format data yang digunakan dalam ilmu atmosfer. Mereka juga tidak menghadapi masalah kecepatan yang saling terkait dan akses yang hampir real-time—dalam perayapan GLEAN, penemuan kumpulan data yang diperbarui dimulai oleh permintaan pengguna manual. Mereka tidak menggunakan katalog OGC atau standar metadata untuk mendukung interoperabilitas.

Proyek BCube (bagian dari inisiatif EarthCube) mengatasi masalah serupa dengan pendekatan lain. EarthCube adalah inisiatif National Science Foundation untuk menciptakan infrastruktur siber berbasis komunitas terbuka bagi semua peneliti dan pendidik di bidang geosains. Infrastruktur siber EarthCube harus mengintegrasikan sumber daya data yang heterogen untuk memungkinkan perkiraan perilaku sistem Bumi yang kompleks. EarthCube terdiri dari banyak blok bangunan. Pekerjaan kami adalah bagian dari blok bangunan EarthCube Cyberway. BCube (The Brokering Building Block) menawarkan pendekatan berbeda untuk interoperabilitas geodata heterogen. BCube mengadopsi kerangka kerja perantara untuk meningkatkan penemuan dan akses data lintas disiplin. Broker adalah layanan data online pihak ketiga yang berisi serangkaian komponen, yang disebut pengakses. Setiap pengakses dirancang untuk berinteraksi dengan jenis repositori geodata yang berbeda. Broker memungkinkan pengguna untuk mengakses beberapa repositori dengan satu antarmuka tanpa memerlukan penyedia data untuk menerapkan langkah-langkah interoperabilitas. BCube mendukung perantara metadata. Itu dapat mencari, mengakses, dan menerjemahkan metadata heterogen dari berbagai sumber. Ini menunjukkan interoperabilitas yang lebih dalam dibandingkan pendekatan lain yang dibahas di sini, namun tidak berupaya memecahkan masalah volume atau kecepatan data. Pendekatan BCube sangat relevan bagi kami; namun, BCube hanya memiliki sedikit dokumen yang tersedia dan sistem tidak dapat diakses. Kami tidak dapat membandingkan beberapa detail dari pendekatan kami yang berbeda.

Song dan Di mempelajari masalah yang sama dengan contoh repositori yang sama: Unidata TDS. Penulis menentukan karakteristik volume dan kecepatan metadata repositori target. Seperti penelitian kami, mereka mengusulkan pemodelannya dengan konsep koleksi dan granul. Mereka menerapkan crawler yang mampu meng-crawl beberapa arsip TDS. Pekerjaan mereka merupakan kemajuan sebelumnya dalam proyek yang sama dengan kami dan sangat relevan dengan penelitian ini. Namun, pendekatan mereka tidak berjalan baik jika menggunakan data TDS dunia nyata, sehingga kami mengambil pendekatan yang berbeda. Kami membangun kembali pekerjaan mereka untuk menunjukkan pencarian real-time dan kemungkinan memproses semua TDS dengan menggunakan model metadata yang lebih canggih, dan klien pencarian terintegrasi yang lebih maju dan layanan pengindeksan yang memungkinkan pencarian real-time yang sebenarnya.

Meninjau pekerjaan yang ada mengungkapkan kemajuan luar biasa dalam memecahkan tantangan dalam menciptakan infrastruktur siber sistem Bumi yang dapat dioperasikan yang secara praktis dapat memproses sejumlah besar dan beragam data observasi dan model yang dihasilkan dalam proses produksi data berkecepatan tinggi. Bidang pekerjaan dalam pemodelan metadata, standardisasi, interoperabilitas, perayapan repositori,

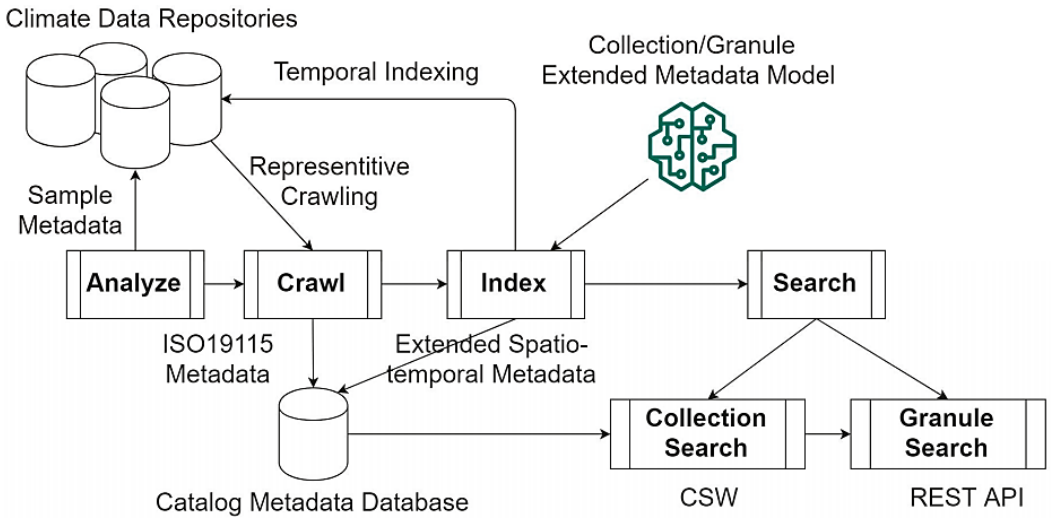
dan pemrosesan memberikan dasar bagi materi penelitian kami. Kontribusi kami adalah mensintesis pendekatan-pendekatan ini untuk mengeksplorasi bagaimana interoperabilitas dan kinerja dapat dicapai secara bersamaan.

11.4 PERANCANGAN MODEL PENCARIAN

Untuk memungkinkan pencarian big data iklim, kami mengusulkan solusi katalog big data baru, yang mencakup langkah-langkah berikut.

1. Menganalisis repositori geodata target yang memberikan contoh yang baik mengenai tantangan data untuk kolaborasi ilmiah sistem Bumi lintas disiplin.
2. Menganalisis kualitas dan karakteristik data dalam repositori yang dipilih.
3. Buatlah model repositori.
4. Gunakan model repositori untuk membangun model sumber daya metadata yang efisien.
5. Mengembangkan sistem perayapan yang menggunakan model sumber daya repositori dan metadata untuk mengoptimalkan algoritme perayapan dan representasi metadatanya.
6. Menunjukkan kemampuan pencarian dan akses geodata besar yang dapat dioperasikan dan dapat dioperasikan yang dimungkinkan oleh pendekatan kami.

Model infrastruktur siber dan arsitektur sistem yang lengkap (berasal dari model metadata kami) ditunjukkan pada Gambar 11.1.



Gambar 11.1. Usulan solusi katalogisasi big data iklim. Singkatan: CSW, Layanan Katalog untuk Web; REST API, Antarmuka Pemrograman Aplikasi Transfer Negara Representasional.

Pemilihan Repositori Metadata

Kami menggunakan Unidata THREDDS Data Server (TDS) sebagai contoh platform repositori geodata target kami. TDS dipilih karena banyak digunakan oleh bidang ilmu atmosfer dan bidang ilmu kebumian terkait lainnya. Ini mendukung beragam metadata terbuka dan standar data dan terdapat banyak pusat data yang menggunakan TDS. Ini mendukung fitur katalog dasar tetapi tidak memiliki kemampuan pencarian lanjutan. Ini

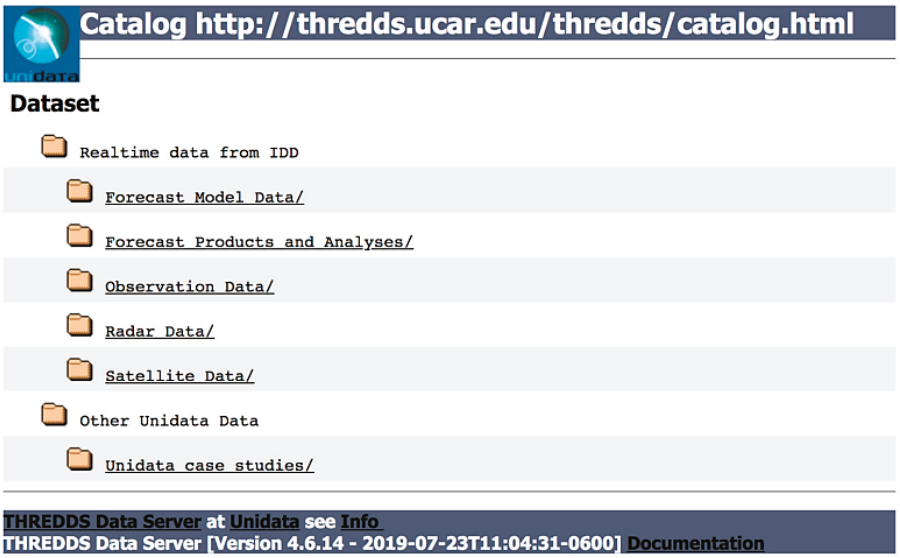
memberi pengguna dan administrator keleluasaan luas tentang bagaimana data diatur dan diperbarui di dalam katalog TDS. Geodata yang disimpan di banyak instalasi TDS memenuhi kriteria luas kami untuk variasi, volume, dan kecepatan data dunia nyata.

Satu instance TDS dipilih sebagai target eksperimen kami. Repositori UCAR Unidata TDS (thredds.ucar.edu) ditentukan sebagai sistem target yang sesuai dan contoh yang baik dari beragam penggunaan TDS. Unidata TDS berisi berbagai data yang diperlukan. Ini memiliki data hampir real-time yang menunjukkan tantangan kecepatan data. Ini berisi berbagai perincian data dan rentang ukuran dan kompleksitas kumpulan data yang tersedia. Volume data dan volume metadata cukup menantang. Struktur katalog bersifat heterogen—berbagai jenis data disusun berdasarkan prinsip berbeda. Pada pemeriksaan awal, Unidata TDS bertekad untuk menjadi contoh bagus dari tantangan yang ingin kami jelajahi.

Dengan menggunakan inspeksi manual dan analisis statistik dasar melalui skrip Python khusus, kami mulai memetakan karakteristik sistem informasi TDS Unidata. Kami mencoba menjawab pertanyaan-pertanyaan berikut: (a) Bagaimana struktur hierarki organisasi data dalam repositori ini?; (b) seberapa sering catatan baru ditambahkan dan dihapus?; (c) bagian mana dari katalog yang menunjukkan pola teratur dalam struktur informasi yang dapat digeneralisasikan dan bagian mana yang berisi informasi unik?; (d) berapa ukuran dan isi sumber metadata yang disimpan dalam katalog?; (e) bagaimana informasi dalam sumber daya metadata terkait dengan lokasi sumber daya metadata dalam hierarki struktur katalog?; dan (f) apa kualitas transmisi data dari sistem jaringan TDS Unidata—bagian informasi TDS apa yang dapat ditransfer dan disalin ke sistem kami?

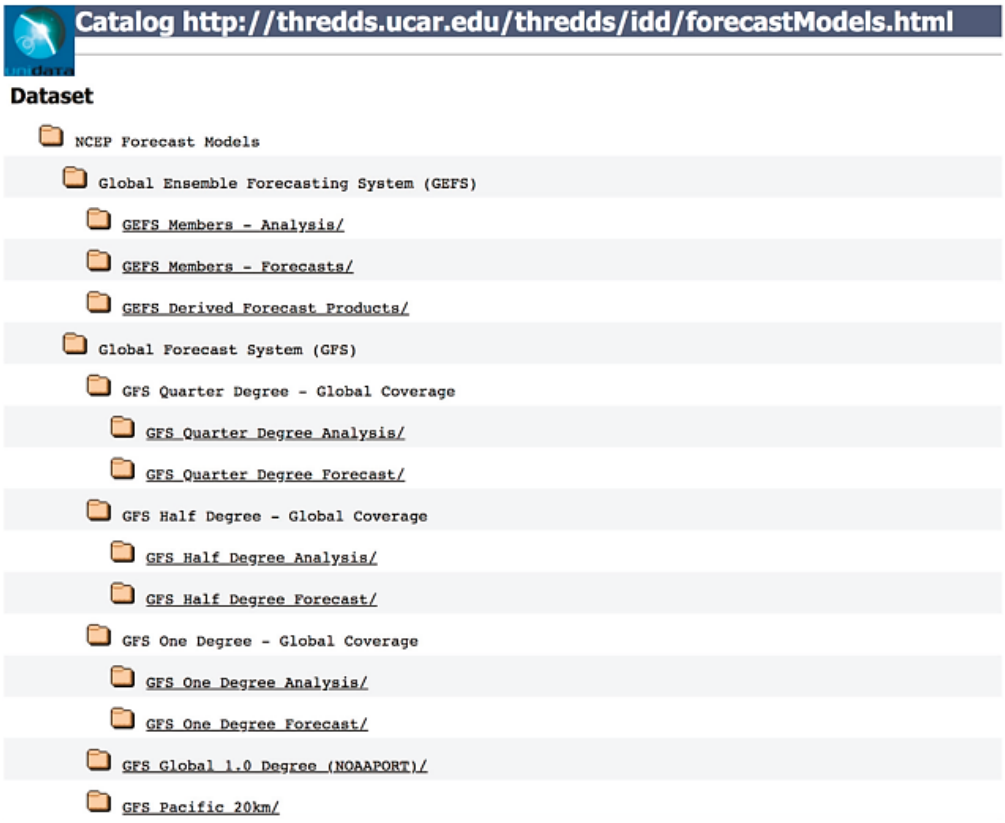
Analisis Repositori

Gambar berikut menunjukkan beberapa struktur permukaan katalog TDS Unidata yang diambil menggunakan browser web dari http://thredds.ucar.edu/thredds/catalog.html. Gambar 11.2 menunjukkan tingkat teratas hierarki katalog. Setiap item yang terdaftar adalah folder (katalog). Kebanyakan katalog berisi beberapa tingkat katalog bersarang (Gambar 11.3) dalam hierarki mirip pohon yang mirip dengan sistem file. Di tingkat pohon (daun) paling bawah (Gambar 11.4), katalog berisi daftar sumber daya data. Katalog disajikan dalam dua format. Pertama adalah format HTML, cocok untuk penjelajahan web manual. Kedua adalah format XML yang berisi metadata tambahan tentang katalog dan sumber data. Representasi XML mengikuti Spesifikasi Katalog Klien THREDDS. Spesifikasi ini memperluas struktur dasar seperti sistem file dengan anotasi metadata deskripsi variabel temporal, spasial, dan data.

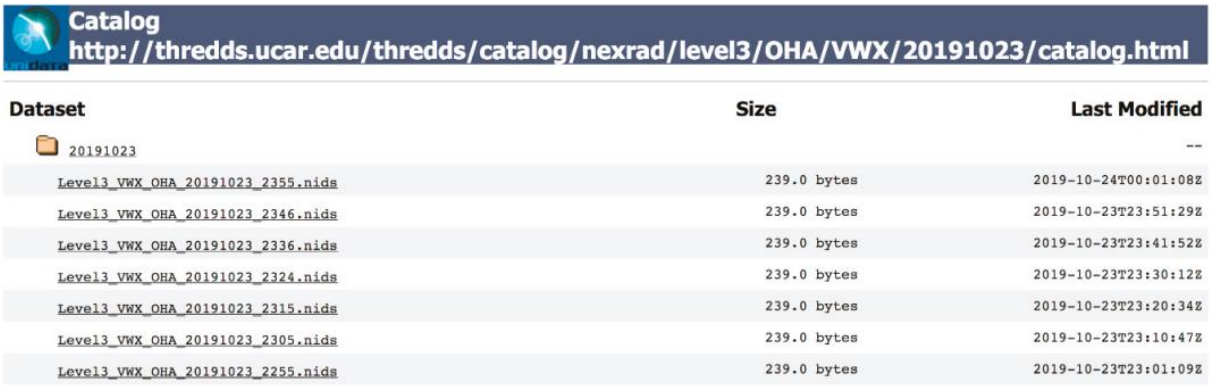


Gambar 11.2. Daftar katalog Unidata THREDDS Data Server (TDS) tingkat atas.

Katalog TDS menyediakan model hierarki katalog umum yang kuat. Namun, penggunaan praktis model ini oleh para ilmuwan yang menghasilkan geodata adalah hal yang menentukan kemungkinan kolaborasi dan harmonisasi data—serta bentuk spesifik dan kemungkinan solusi untuk permasalahan big data. Korespondensi email dengan Unidata menjelaskan bahwa data yang ditempatkan di sub-katalog berbeda diproduksi dan diatur oleh tim ilmuwan berbeda. Meskipun Unidata TDS bertindak sebagai gudang terpadu untuk beragam data Bumi, tidak ada prinsip pengorganisasian wajib yang menyeluruh untuk memungkinkan harmonisasi data.



Gambar 11.3. Katalog bersarang. Hanya 11 entri pertama yang ditampilkan di sini. Entri lainnya dihilangkan.



Gambar 11.4. Daftar sumber daya data (kumpulan data) di bagian bawah hierarki katalog. Hanya tujuh entri pertama yang ditampilkan di sini. Entri lainnya dihilangkan.

Oleh karena itu, langkah selanjutnya adalah memahami dan mendeskripsikan sub-struktur berbeda yang secara organik diadopsi oleh tim berbeda. Setelah inspeksi manual dan analisis statistik dasar dilakukan dengan skrip Python khusus, informasi berikut dikumpulkan untuk menjelaskan secara luas berbagai pola pemanfaatan sub-katalog (Tabel 11.1).

Tabel 11.1. Karakteristik big data subkatalog Unidata TDS.

Katalog	SubKategori/ Granule	Estimasi Ukuran Katalog	Produksi Baru Katalog	Granulity	Regularity
NCEP Forecasy	5300/5300	500 MB	6 Jam	Kasar	Regular
Observation	8/186	20 MB	Irregular	Kasar	Irregular
Satelit	1.500/7.100	150 MB	10 Menit	Baik	Regular
Radar	25.000/7Juta	25 GB	5-10 Menit (irregular)	Sangat Baik	Irregular

Empat jenis data umum disimpan secara bersamaan di repositori TDS Unidata: (1) keluaran model prakiraan, (2) observasi (rangkaian waktu dari instrumen in-situ), (3) citra satelit, dan (4) citra radar dari radar stasioner. jaringan (NEXRAD, Radar Cuaca Generasi Selanjutnya). Setiap jenis berisi lebih banyak variasi tambahan dalam hierarki subkatalognya masing-masing, namun pada tingkat ini, terdapat beberapa perbedaan luas yang jelas dan berguna dalam kualitas data yang dapat memandu eksperimen kami.

Pada Tabel 11.1, perkiraan ukuran katalog adalah ukuran total metadata yang disimpan dalam katalog. Sebagian besar metadata ini benar-benar mubazir, namun tanpa mengetahui struktur yang lebih dalam dari data ini, kami harus mencerminkan semua data ini untuk mengaktifkan kemampuan pencarian dan penemuan yang tidak didukung oleh THREDDS. Kami menghitung throughput transfer data maksimum sebesar 4 MB/s atau 5 menit untuk memuat 1 GB data katalog. Tampaknya mungkin untuk mencerminkan seluruh katalog metadata TDS Unidata dalam beberapa jam, namun throughput data yang kami amati tidak konsisten, sering kali melambat satu urutan besarnya. Selain itu, kecepatan pemrosesan data (pengindeksan dan pendaftaran dengan katalog OGC CSW yang sesuai standar) juga sangat memakan waktu, serta sumber daya komputasi dan penyimpanan. Kami tidak memiliki kemampuan untuk mendaftar dan mencari jutaan catatan yang sebagian besar berisi informasi berlebihan. Selain itu, data TDS Unidata ditambahkan hampir secara real-time sesuai dengan pola dan struktur tertentu dalam sub-katalog. Jika kami mencoba menyalin dan mendaftarkan semua metadata tersebut, kami tidak akan mampu menyediakan kemampuan yang hampir real-time.

Dua kolom terakhir pada Tabel 11.1 menunjukkan dua kualitas penting yang menentukan pendekatan apa yang perlu kami ambil untuk mengintegrasikan metadata tersebut ke dalam sistem kami.

Jika kumpulan data akhir memiliki perincian “kasar”, artinya setiap kumpulan data adalah sebuah file dan ukuran metadatanya kecil jika dibandingkan dengan ukuran datanya — untuk kumpulan data “kasar”, kita dapat menyalin, memanen, dan mengindeks metadata tersebut ke dalam penelusuran kita sistem. Kumpulan data “Baik” memperluas kemampuan teknis untuk mentransfer dan memproses metadata. Catatan yang “sangat bagus” terlalu banyak (file datanya terlalu kecil) sehingga kami tidak dapat menyinkronkan atau memproses metadatanya secara efektif.

Jika kumpulan data diproduksi dengan cara biasa (atribut spatiotemporal yang dapat diprediksi), maka kita dapat mengumpulkan informasi minimal dan memodelkan seluruh katalog. Namun, untuk metadata radar NEXRAD, tidak ada pola reguler dalam produksi metadata. Catatan baru dapat ditambahkan setiap 5 menit atau setiap 15 menit—dan keteraturan/ketidakteraturannya juga bervariasi dari waktu ke waktu dan bergantung pada lokasi radar yang berbeda (sub-katalog berbeda). Data tidak beraturan yang terperinci ini adalah yang paling menantang, karena data tersebut tidak dapat diambil secara besar-besaran atau dimodelkan dengan cara yang akurat. Hal ini memerlukan pendekatan kombinasi yang ditargetkan. Pertimbangan tambahan muncul ketika melacak kumpulan data apa yang telah kedaluwarsa dan dihapus—idealnya, hal ini harus dilakukan tanpa melakukan pemindaian penuh yang mahal terhadap repositori TDS.

Pemeriksaan lebih lanjut terhadap struktur sub-katalog untuk data yang sangat granular tidak beraturan (dan teratur) mengungkapkan informasi struktural tambahan yang berguna. Beberapa katalog bersifat "dinamis" (atau "langsung" atau "streaming")—katalog tersebut diperbarui dengan sumber data baru dengan frekuensi reguler (atau tidak teratur). Katalog lain bersifat arsip—dapat diasumsikan tidak pernah berubah (sampai habis masa berlakunya dan dihapus seluruhnya). Tiga jenis sub-katalog yang berbeda dapat diidentifikasi:

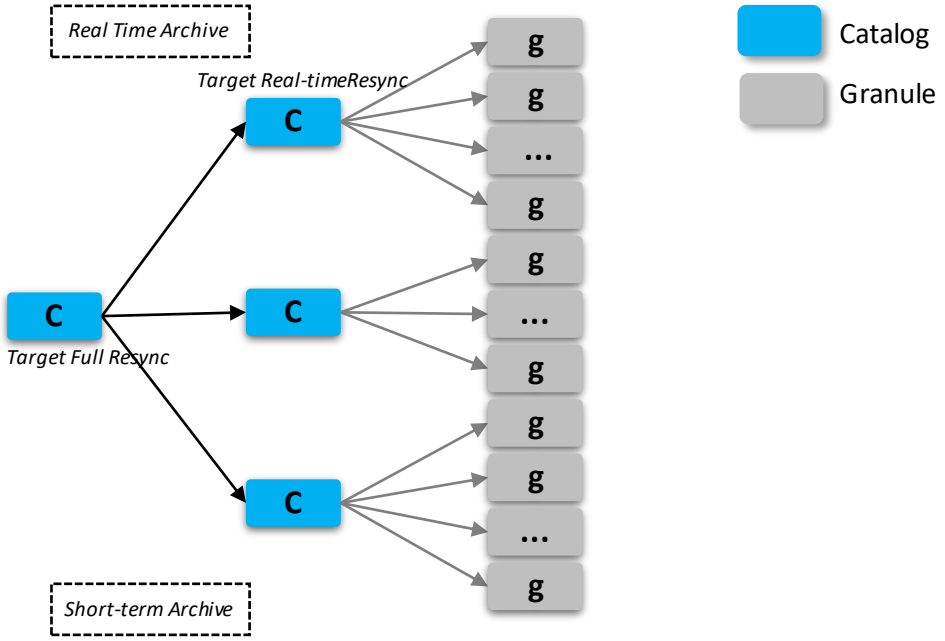
- *Direktori arsip murni:* Folder ini hanya berisi koleksi dan butiran lama dan tidak akan pernah diperbarui atau dihapus.
- *Direktori arsip campuran:* Beberapa sub-folder berisi materi arsip, beberapa berisi streaming langsung dan kumpulan data hampir real-time.
- *Direktori arsip harian:* Folder yang berisi data streaming untuk hari tertentu; ketika hari berlalu, direktori ini menjadi folder arsip dan tidak perlu dicerminkan lagi. Ketika arsip harian habis masa berlakunya, semua sumber data untuk hari itu akan dihapus bersama-sama.

Katalog data besar biasanya perlu menyelesaikan banyak tugas perayapan untuk mengambil file metadata, dan mengulangi pemindaian untuk menangkap metadata dari kumpulan data yang baru diamati secara rutin. Perayapan adalah sumber informasi mendasar dari metadata, dan cara merayapi secara cerdas adalah salah satu tantangan terbesar dalam penelusuran data besar karena beban komputasi yang berulang dan kompleksitas konten. Saat merancang strategi perayapan, kami mempertimbangkan frekuensi pembaruan pengamatan, jangka waktu, organisasi jaringan observatorium, dan membuat perayap hanya menyentuh folder sensor yang diperbarui tersebut pada tingkat pengumpulan (sensor). Meskipun sebuah sensor memiliki jutaan catatan metadata, kami hanya meng-crawl metadata tersebut pada tingkat sensor. Dengan kata lain, hanya satu metadata yang dirayapi untuk setiap sensor (atau instrumen). Dengan menggunakan strategi ini, kami dapat menghemat banyak waktu dalam perayapan dan transfer metadata melalui jaringan, terutama saat jaringan tidak stabil. Setelah menerapkan mekanisme pekerja paralel, kami dapat memiliki lusinan crawler yang bekerja untuk memindai dan menangkap metadata baru/yang diperbarui sebesar petabyte kumpulan data iklim.

Perayap kami berbeda dari kebanyakan perayap yang ada dalam literatur, karena ini bukan perayap mesin pencari untuk tujuan umum. Perayap pada umumnya mengunduh seluruh halaman web, menemukan tautan untuk diikuti, dan menambahkan tautan tersebut ke antrean pekerjaan. Kami tidak dapat melakukan hal serupa, karena konten web yang kami jelajahi (katalog TDS) berisi informasi yang sangat berlebihan sehingga tidak mungkin diunduh dan diproses secara keseluruhan tanpa membebani sumber daya komputasi dan jaringan yang tersedia. Ada berbagai sensor dalam jaringan pemantauan iklim dan sensor tersebut berubah secara dinamis, dengan sensor baru ditambahkan atau sensor lama dihilangkan. Kami harus merayapi Server Data THREDDS untuk memastikan semua pengamatan disinkronkan sepenuhnya dalam katalog kami. Desain perayap kami harus menggabungkan pengetahuan tentang metadata dan struktur metadata, algoritma pemrosesan dan antriannya agar hanya mengunduh informasi penting.

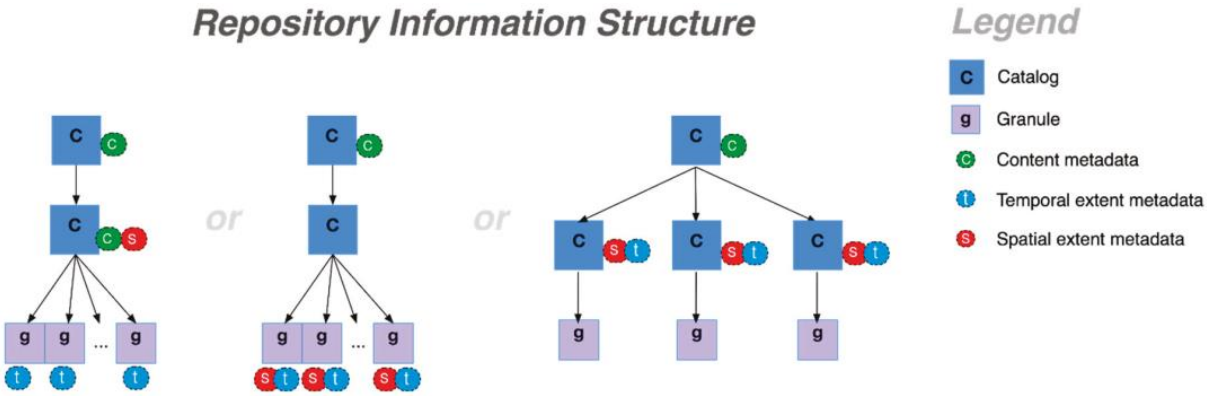
Pengindeksan

Langkah ketiga adalah pengindeksan, yang mengekstrak informasi spatiotemporal dari metadata yang dirayapi dan membuat indeks untuk butiran data deret waktu oleh setiap instrumen. CSW menyediakan registrasi metadata dasar dan model kueri. Namun, besarnya rincian objek metadata (dan kurangnya kemampuan agregasi/relasional) membuat CSW tidak efisien dalam menyimpan dan menanyakan kumpulan data dalam jumlah besar dan hanya memiliki sedikit variasi dalam metadatanya. Diperlukan model yang lebih efisien. Ini adalah masalah yang telah lama dieksplorasi dan pada dasarnya telah dipecahkan dalam ilmu komputer dan informatika. Theodoridis dkk. merangkum pendekatan dasar. Untuk objek spatiotemporal yang berevolusi terhadap waktu, cuplikan evolusinya dapat diwakili oleh triplet {o_id, si, ti}—id objek, stempel ruang, dan stempel waktu. Informasi ini memungkinkan kami membuat “model produksi repositori” (Gambar 11.5). Kami mengidentifikasi pola dalam struktur hierarki katalog yang memungkinkan kami mengidentifikasi jalur mana dalam hierarki folder katalog yang “aktif” dan mana yang “arsip”. Dalam implementasi crawler kami (dibahas di bagian selanjutnya), kami menggunakan pola jalur struktur untuk menggerakkan algoritme crawler dalam dua tahap—tahap “sinkronisasi penuh”, yang menyalin data arsip, dan tahap “perbarui”, yang memantau dan menyegarkan data daftar dari jalur katalog "langsung".



Gambar 11.5. Model produksi metadata repositori.

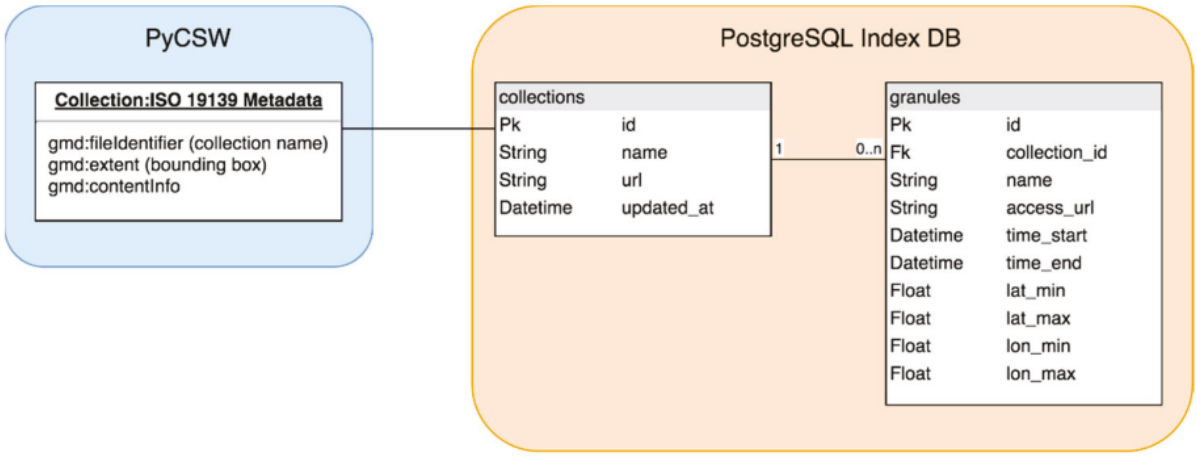
Model produksi repositori memungkinkan perayapan yang ditargetkan—namun, jumlah sumber daya metadata masih terlalu besar untuk diambil, diproses, dan diindeks secara keseluruhan, bahkan ketika dilakukan dalam dua tahap untuk menghindari pengambilan yang berlebihan. Kami membutuhkan model kedua yang mencakup struktur informasi metadata (Gambar 11.6). Ada dua masalah yang perlu kami selesaikan: Pertama, sebagian besar metadata dalam katalog benar-benar mubazir; kedua adalah cakupan informasi metadata tidak konsisten dalam katalog. Kedua masalah tersebut memiliki sumber yang sama: Katalog, dan subkatalog serta butiran data, semuanya dapat memiliki metadata yang dilampirkan.



Gambar 11.6. Model informasi repositori.

Dalam contoh dari Unidata TDS ini, kita melihat bahwa metadata dilampirkan ke struktur katalog hierarki dengan berbagai cara. Pada contoh pertama, katalog berisi beberapa metadata konten (misalnya: Kepengarangan), subkatalog berisi metadata konten tambahan (misal: Nama variabel) dan metadata spasial, sedangkan setiap butiran berisi metadata

temporal. Dalam dua contoh berikutnya, distribusi metadata antara katalog dan butiran berbeda. Contoh terakhir adalah kasus dimana setiap katalog hanya berisi satu record data (butiran). Dalam beberapa kasus, metadata hanya diduplikasi di antara beberapa tingkat katalog, sementara di kasus lain, satu lapisan tertentu berisi semua metadata. Detail penting lainnya adalah hierarki katalog, nama katalog induk juga merupakan metadata untuk sumber data. Jika digabungkan, kedua perspektif ini (model perubahan informasi dan model struktur informasi) menghasilkan model repositori TDS Unidata yang dapat digunakan untuk mengembangkan pengumpulan dan representasi semua metadata yang ada secara efisien (tidak berlebihan). Dengan menerapkan model produksi pada desain crawler, kami hanya dapat mengumpulkan informasi yang kami tahu telah berubah. Mengetahui struktur perubahan data juga memungkinkan kami melakukan pengumpulan bertahap yang ditargetkan untuk kemampuan penemuan hampir secara real-time. Kami mendefinisikan dua jenis objek: Koleksi dan butiran. Koleksi berisi metadata konten (judul, deskripsi, kepengarangan, informasi variabel/band, dll.). Setiap koleksi berisi satu atau lebih butiran. Setiap butiran hanya berisi metadata tingkat spatiotemporal. Standar katalog OGC CSW tidak mendukung komposisi koleksi dan butiran, sehingga kami menggunakan CSW untuk mewakili koleksi saja, sedangkan butiran harus disimpan secara eksternal. Kami menggunakan perangkat lunak PyCSW populer untuk menyimpan metadata koleksi. Kami memperluas PyCSW dengan database relasional PostgreSQL untuk menyimpan hubungan antara koleksi dan butiran serta metadata butiran (Gambar 11.7).

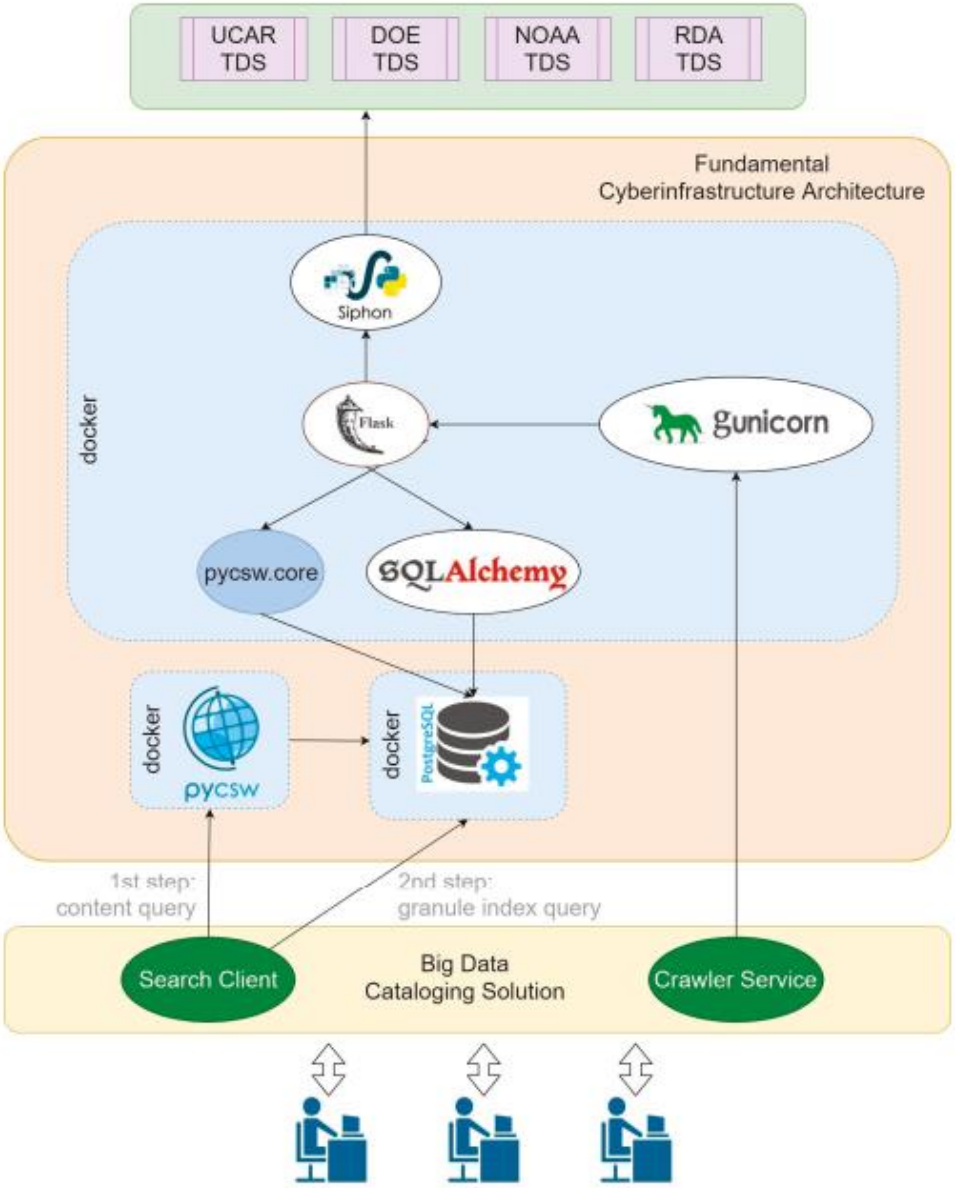


Gambar 11.7. Pengumpulan metadata dan sumber daya granula yang disimpan dalam database PyCSW dan PostgreSQL yang tertaut secara referensial. PyCSW gmd: fileIdentifier sesuai sebagai kunci untuk kolom nama tabel koleksi di database SQL

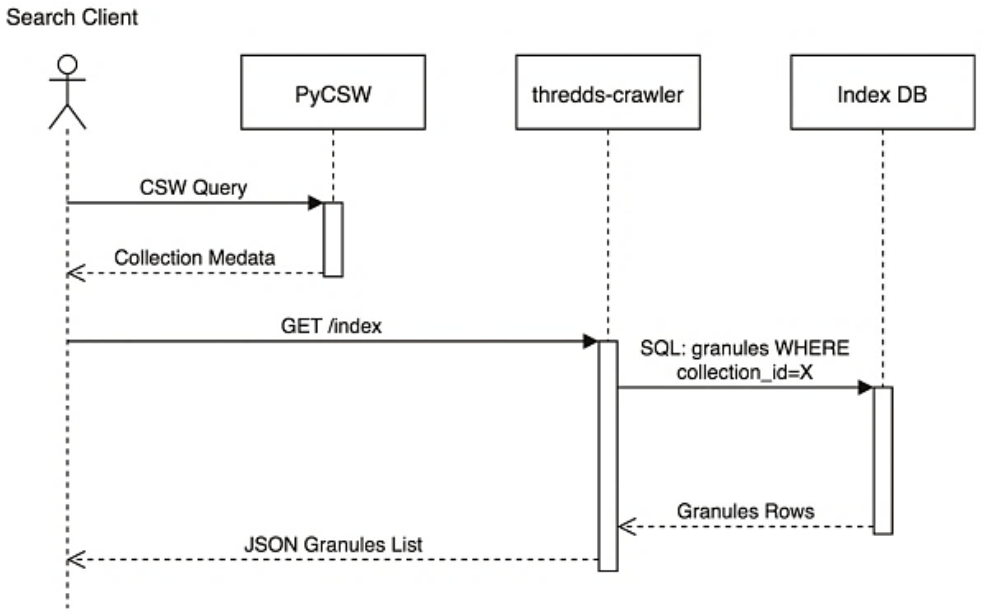
Proses Pencarian Dua Langkah

Ketika metadata dikumpulkan ke dalam PyCSW dan indeks butiran temporal disimpan di PostgreSQL, klien pencarian dapat menggunakan dua sumber data ini untuk mengambil hasil akhir untuk diakses.

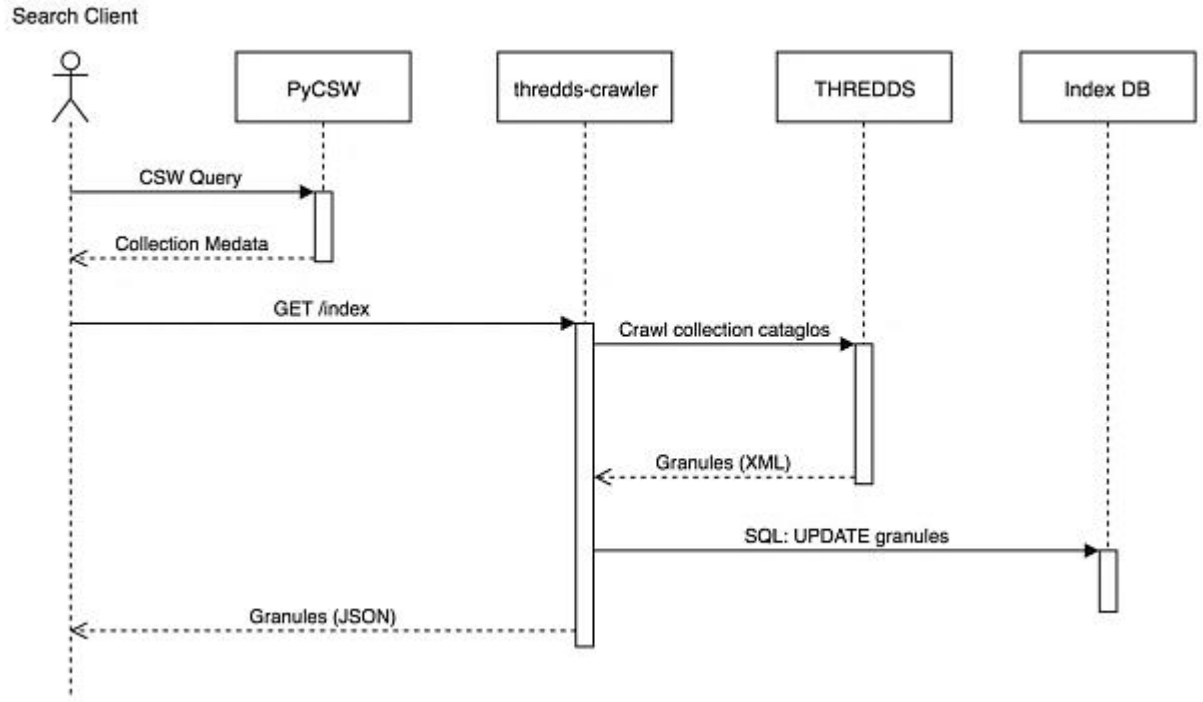
.



Gambar 11.8. Arsitektur implementasi pencarian big data yang disajikan melalui Unidata THREDDS Data Server (TDS). Meskipun dalam penelitian kami hanya UCAR TDS yang digunakan, sistem ini dirancang untuk mendukung repositori TDS apa pun sebagai sumber data.



Gambar 11.9. Pengambilan indeks granula sederhana selama pencarian.



Gambar 11.10. Pengambilan indeks granul ketika rentang temporal granul berada di luar rentang yang disimpan dalam indeks. Pencarian memicu langkah tambahan yang segera meng-crawl katalog TDS dan memperbarui metadata granul secara real-time. Singkatan: DB, Basis Data.

Proses pencarian berlangsung dalam dua langkah. Awalnya, klien mencari toko PyCSW menggunakan metode pencarian dan kueri standar. Ini mengembalikan daftar hasil tingkat pengumpulan. Untuk mendapatkan daftar butiran, klien pencarian mengirimkan kueri kedua ke layanan perayap. Layanan perayap menanyakan indeks butiran, menyegarkan indeks

dengan butiran terbaru jika diperlukan, dan mengembalikan daftar butiran untuk koleksi yang diminta. Klien pencarian kemudian dapat menggunakan catatan CSW tingkat kumpulan dan menggabungkannya dengan informasi butiran yang dipilih untuk menghasilkan informasi CSW tingkat butiran. Gambar 11.1 dan 11.8–10 menunjukkan interaksi ini dari perspektif arsitektur sistem dan rangkaian peristiwa.

Sejauh ini, kami telah menganalisis struktur repositori Unidata TDS, membangun model repositori yang dapat menginformasikan strategi perayapan yang efektif, dan menentukan model keluaran produk untuk crawler. Kami juga telah menjelaskan bagaimana seharusnya klien pencarian berfungsi. Untuk menyelesaikan percobaan kami, kami membuat perayap yang mengikuti model metadata kami dan menunjukkan kemampuan penelusuran web untuk seluruh konten Unidata TDS.

11.5 IMPLEMENTASI

Kami menerapkan sistem modul dalam EarthCube CyberConnector untuk mewujudkan mode yang diusulkan (Gambar 11.8). Implementasinya meliputi sistem pencarian server dan sistem klien. Kami akan memperkenalkan kemampuan pencarian yang diaktifkan oleh sistem ini.

Implementasi Layanan Perayap

Kami membangun perayap web yang melintasi Unidata TDS dan mengekstrak serta menyimpan metadata penting tanpa menggunakan sumber daya yang tidak perlu. Namanya 'thredds-crawler' dan kode sumbernya tersedia melalui repositori GitHub publik: https://github.com/CSISS/thredds-crawler.

Perayap ditulis dengan Python. Itu dibangun menggunakan perpustakaan sumber terbuka umum untuk interaksi HTTP API (Flask [https://www.fullstackpython.com/flask.html], Gunicorn [https://gunicorn.org/]), pemrosesan XML umum (libxml [https ://lxml.de/}]), dan abstraksi database (SQLAlchemy [https://www.sqlalchemy.org/]). Ia menggunakan pustaka threading Python asli untuk mendukung konkurensi. Untuk melintasi katalog Unidata THREDDS dan mengambil metadata, ia menggunakan pustaka Python Siphon yang disediakan Unidata [https://github.com/Unidata/siphon].

Untuk mendukung persyaratan eksperimen big data kami, crawler terintegrasi erat dengan perangkat lunak katalog PyCSW [https://pycsw.org/] dan database PostgreSQL [https://www.postgresql.org/]. Crawler, PyCSW, dan database masing-masing berjalan di container Docker [https://www.docker.com/] yang terpisah. Demi demonstrasi ini, ketiga layanan berjalan pada mesin yang sama dan berkomunikasi melalui jaringan lokal. Alat penulisan Docker digunakan untuk menghubungkan dan mengatur ketiga container. Arsitektur ini memungkinkan penskalaan sederhana ke beberapa mesin menggunakan container, sehingga memungkinkan potensi peningkatan substansial dalam kinerja sistem.

Kontainer buruh pelabuhan perayap berjalan sebagai layanan web yang dihosting oleh Gunicorn—server HTTP python yang banyak digunakan untuk menghosting aplikasi web. Ini melayani tiga titik akhir HTTP API yang menjalankan fungsi berikut: Memanen, membuat indeks, dan membaca indeks.

Fungsi panen memuat XML Katalog Unidata dari catalog_url yang ditentukan menggunakan perpustakaan Siphon. Katalog berisi daftar kumpulan data. TDS memiliki fitur untuk menerjemahkan metadata kumpulan datanya ke dalam format XML yang kompatibel dengan ISO/OGC. Untuk setiap kumpulan data yang diambil, pemanen membuat kueri ke TDS untuk mengambil metadata ISO/OGC untuk kumpulan data tersebut. Namun, metadata ISO/OGC yang dikembalikan oleh TDS seringkali tidak lengkap, tidak akurat, atau tidak konsisten. Proses pemanenan perayap kemudian menerapkan serangkaian filter XML ke metadata ISO/OGC untuk memperbaikinya dengan informasi dari metadata kumpulan data TDS asli. Setelah metadata diunduh dan diproses, metadata disimpan langsung di database PyCSW menggunakan pustaka kompatibilitas PyCSW.

Pengindeksan mirip dengan pemanenan, namun melibatkan strategi untuk menargetkan kumpulan data yang akan dipanen dan langkah pemrosesan tambahan. Selama pembuatan indeks, katalog dan kumpulan data TDS diubah menjadi koleksi dan butiran dalam model kami. Untuk setiap dataset TDS yang ditemukan, kami menentukan nama koleksinya. Yang terpenting, nama koleksi bukanlah nama katalog yang berisi dataset tersebut. Kami menemukan bahwa nama katalog tidak konsisten, namun id kumpulan data TDS berisi informasi identifikasi yang konsisten. Di TDS, id kumpulan data dibuat unik dengan menyertakan stempel waktu dalam id kumpulan data. Misalnya, dalam kumpulan data dengan id "NWS/NEXRAD3/PTA/YUX/20190830/Level3_YUX_PTA_20190830_1713.nids", bagian "20190830_1713" adalah stempel waktu. Untuk mengubah katalog TDS dengan kumpulan data menjadi koleksi dengan butiran, kami menghapus informasi temporal untuk membuat id koleksi. Kemudian, kami mengunduh kumpulan data dalam format ISO/OGC XML dan mengubah konten XML-nya dengan fungsi filter yang disebut "pembuat koleksi". Fungsi ini memperbarui metadata kumpulan data untuk mengubahnya menjadi bentuk yang lebih umum yang menjelaskan koleksi. Ini mengubah pengidentifikasi yang disimpan dalam metadata. Ini juga menambahkan bidang tambahan yang memenuhi standar yang mengidentifikasi metadata untuk mendeskripsikan "seri" ("seri" dalam model ISO/OGC, "koleksi" dalam model kami). Proses ini perlu dilakukan hanya untuk kumpulan data pertama yang ditemukan pada setiap koleksi. Saat memproses kumpulan data tambahan, koleksi yang ada digunakan kembali. Di TDS, informasi tingkat spatiotemporal kumpulan data adalah bagian dari metadata katalog, yang berarti kita hanya perlu mengunduh satu metadata kumpulan data untuk membangun metadata koleksi dan kita dapat mengindeks sisa butiran dari metadata katalog. Ini memecahkan masalah redundansi yang sebelumnya menghalangi pencarian TDS. Kami juga mengoreksi pengidentifikasi TDS untuk memastikan bahwa bagian otoritas namespace dari pengidentifikasi tersebut disetel dengan benar.

Tabel berikut membantu mengilustrasikan proses mengekstrak koleksi dari pengidentifikasi granul untuk beberapa tipe data. Tabel 11.2 menunjukkan jalur katalog kumpulan data TDS. Tabel 11.3 menunjukkan bagaimana pengidentifikasi koleksi dihasilkan, dan Tabel 11.4 menunjukkan nama koleksi hasil akhir.

Tabel 11.2. Jalur katalog kumpulan data TDS untuk tiga jenis data. Hierarki jalur katalog ditandai dengan warna hijau. Nama file kumpulan data ditandai dengan warna merah.

Tipe Data	Contoh Path Katalog TDS
RADAR	Radar Data, NEXRAD Level iii Radar, YUX, 20190830, **Level3_YUX_PTA_20190830_1713.nids**
Model	Forecast Model data, GEFS Member, Analisis GEFS_Global_1p0deg_Ensemble_ana_20190731_0000.grib2, **GEFS_Global_1p0deg_Ensemble_ana_20190731_0000.grib2,**
Satelit	**Satellite Data, GOES West Product, CloudAndMoistureImagery, Mesoscale-2,** Channel16,20190831, **OR_ABI-L2-CMIPM2-M6C16_G17_s20192430003570_e20192430003570_c20192430003570.nc**

Tabel 11.3. Pengidentifikasi kumpulan data TDS untuk tiga jenis data. Bagian pengidentifikasi yang berisi informasi temporal disorot.

Tipe Data	Contoh Identifier Dataset TDS
RADAR	NWS/NEXRAD3/PTA/YUX/20190830/Lelev3_YUX_PTA_20190830_1713.nids
Model	Grib/NCEP/GEFS/Global_1p0deg_Ensemble/Member-analysis/ GEFS_Global1p0deg_Ensemble_ana_2019731_000.grib2 goes-west-products/CloudAndMoistureImagenery/Mesoscale-2/Channel16/16/20190831/OR_ABI-I.2-CMIPM2-M6C16_G17_s20192430003570_e20192430003570_c20192430003570.nc
Satelit	

Tabel 11.4. Pengidentifikasi kumpulan untuk contoh ID kumpulan data. Mereka dihitung dengan menghapus informasi temporal dan mengawali bidang namespace otoritas.

Tipe Data	Contoh Pengidentifikasi Kumpulan data terhitung
RADAR	edu.ucar.unidata:NWS/NEXRAD3/PTA/YUX/Level3_YUX_PTA.nids
Model	edu.ucar.unidata:grib/NCEP/GEFS/Global_1p0deg__nsemble/member-analysis/GEFS_Global_1p0deg_Ensemble_ana.grib2
Satelit	edu.ucar.unidata:goes-west-products/CloudAndMoistureImagery/Mesoscale-2/Channel16/OR_ABI-I.2-CMIPM2-M6C16_G17.nc

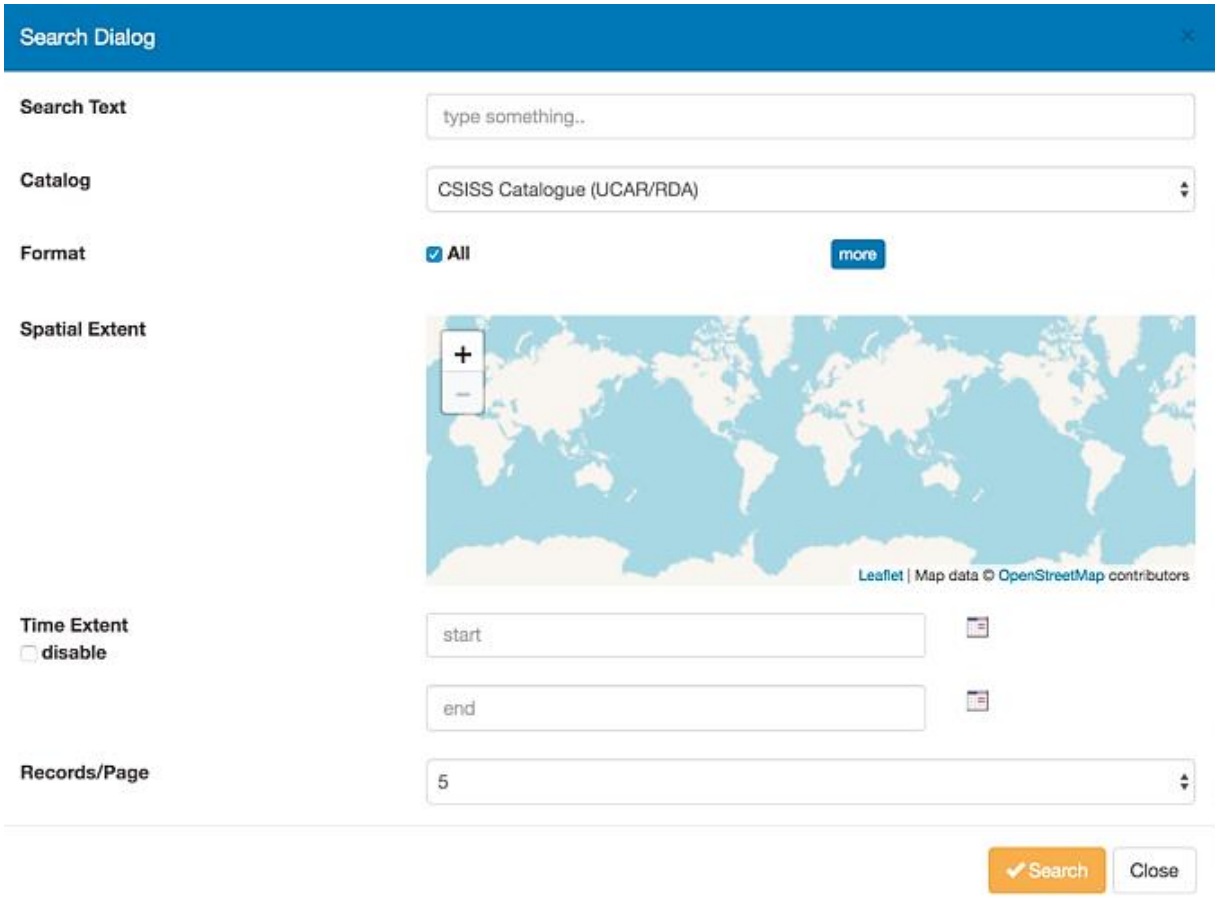
Ketika pengumpulan indeks selesai, informasi pengumpulan (format metadata XML OGC/ISO 19139) disimpan di PyCSW. Informasi granula disimpan dalam penyimpanan indeks SQL yang ringkas (Gambar 11.7). Setelah indeks dibuat, indeks dapat diambil dari layanan web crawler menggunakan HTTP API (GET/index). Permintaan ini menggunakan nama koleksi dan jangkauan waktu sebagai parameter. Meskipun model data kami mencakup tingkat spasial dan temporal yang terperinci, pada saat publikasi, hanya kueri indeks temporal yang diterapkan. Ia memeriksa penyimpanan data indeks untuk melihat apakah butiran terbaru yang tersedia lebih baru dari jangka waktu yang diminta. Jika butiran yang lebih baru tidak diperlukan, crawler akan mengembalikan daftar butiran dalam format JSON yang ringkas (Gambar 11.9). Namun, jika indeks tidak berisi butiran yang cukup baru, maka layanan indeks akan melakukan pengindeksan "penyegaran" sebagian pada repositori TDS. Ia menggunakan tautan katalog TDS yang disimpan dalam koleksi PyCSW kami dan menjalankan kembali proses

indeks yang dijelaskan di sini. (Gambar 11.10). Namun, seperti yang kita diskusikan di bagian Eksperimen, katalog TDS disusun dengan beberapa sub-katalog yang menyimpan informasi arsip, sementara sub-katalog lainnya berisi "data langsung" yang hampir real-time. Proses penyegaran indeks crawler memanfaatkan struktur tersebut. Ini mengabaikan sub-katalog lama dan hanya mengindeks sub-katalog yang berisi data lebih baru dan tidak diketahui. Hal ini membuat pengambilan indeks hampir real-time menjadi cepat dan efisien.

Pemanenan dan pembuatan indeks menggunakan strategi antrean multi-utas yang sama untuk mencapai kinerja yang lebih tinggi. Biasanya, sebagian besar waktu dihabiskan menunggu data dikirim melalui jaringan. Dengan menggunakan banyak thread, kita dapat meningkatkan saturasi jaringan dan komputer lokal serta sumber daya memori, yang memungkinkan metadata tersedia lebih cepat.

Implementasi Sistem Pencarian

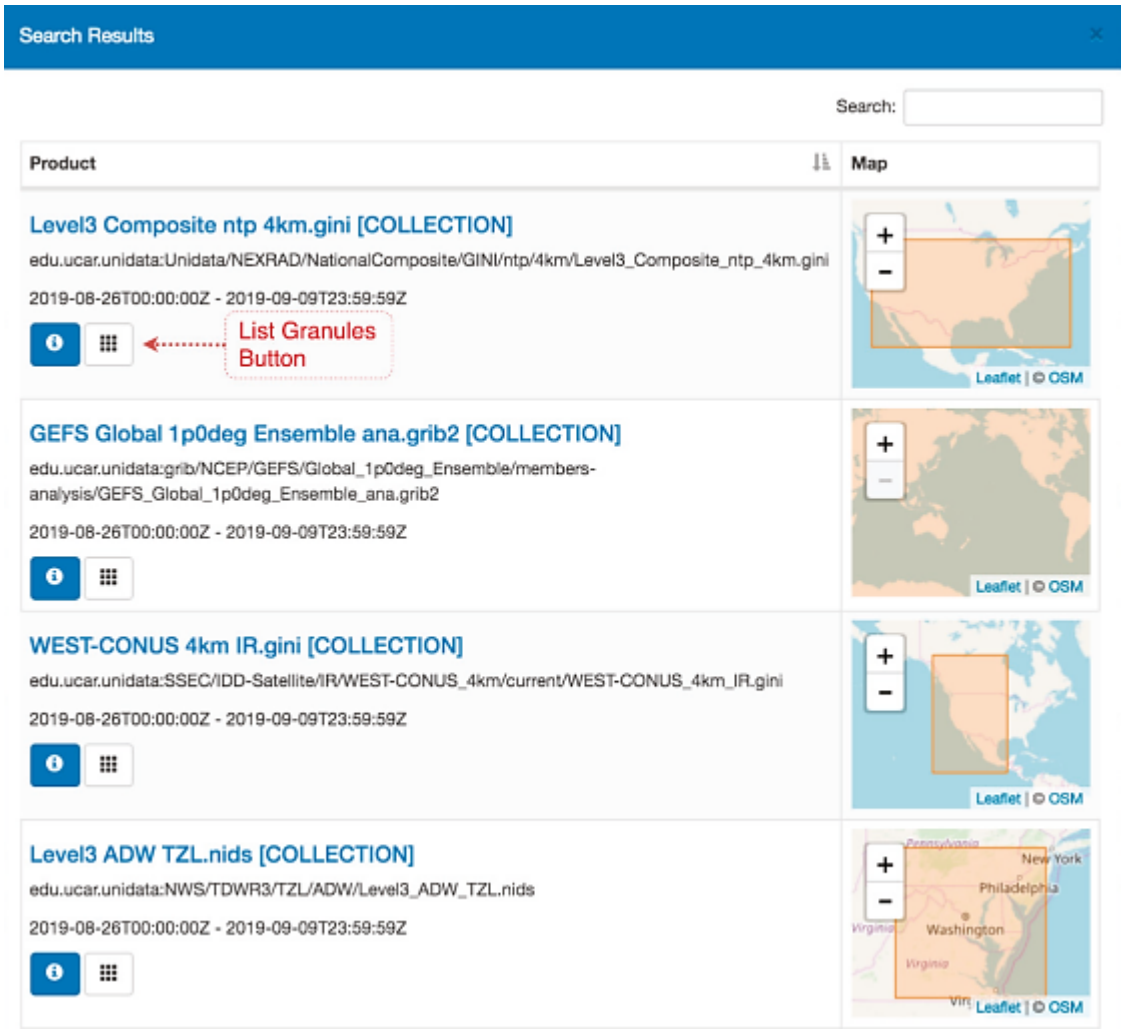
Sistem pencarian diimplementasikan berdasarkan blok bangunan infrastruktur EarthCuber CyberConnector yang dikembangkan sebelumnya.



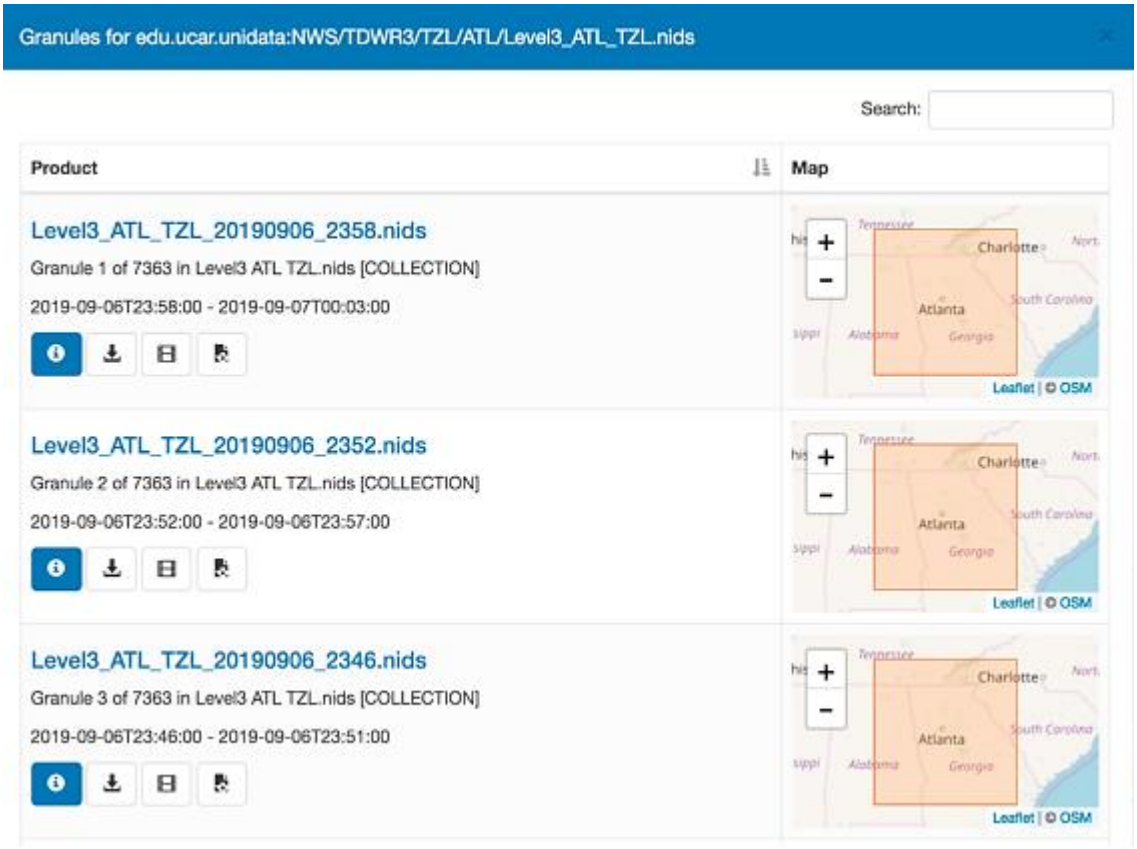
Gambar 11.11 Antarmuka web klien pencarian.

CyberConnector adalah aplikasi web berbasis Java yang mendukung penemuan dan visualisasi data dari katalog CSW. Kami memperluas CyberConnector untuk mendukung akses metadata yang dikumpulkan dan diindeks oleh perayap thredds yang dijelaskan di bagian sebelumnya. Kami memodifikasi Klien Pencarian CyberConnector untuk melakukan pencarian

dua tahap. Pengguna aplikasi web memilih fungsi “Search” (Gambar 11.11). Mereka memilih rentang waktu, yang digunakan oleh layanan indeks perayap thredds untuk menentukan apakah penyegaran granula diperlukan. Browser web mengirimkan permintaan AJAX ke aplikasi web CyberConnector dengan parameter pencarian. CyberConnector menanyakan layanan PyCSW perayap thredds untuk koleksi yang cocok dengan parameter kueri. Ini mengembalikan daftar koleksi. Untuk melihat butiran yang tersedia dalam koleksi, pengguna mengklik tombol “Daftar Butiran” (Gambar 11.12). Ini mengeluarkan permintaan lain ke CyberConnector untuk daftar butiran dalam batas waktu yang ditentukan. Aplikasi web CyberConnector memproksi permintaan daftar butiran ke layanan pengindeksan perayap thredds, yang mengembalikan daftar butiran (Gambar 11.9); atau thredds-crawler memanen TDS untuk memperbarui indeks dan kemudian mengembalikan daftar butiran (Gambar 11.10). Klien menerima daftar butiran, yang kemudian dapat diunduh atau divisualisasikan (Gambar 11.13).



Gambar 11.12. Hasil pencarian dengan tombol “List Granules”.



Gambar 11.13. Daftar butiran untuk koleksi. Tombol tersebut memungkinkan pengguna melihat metadata, mengunduh kumpulan data, atau memvisualisasikannya.

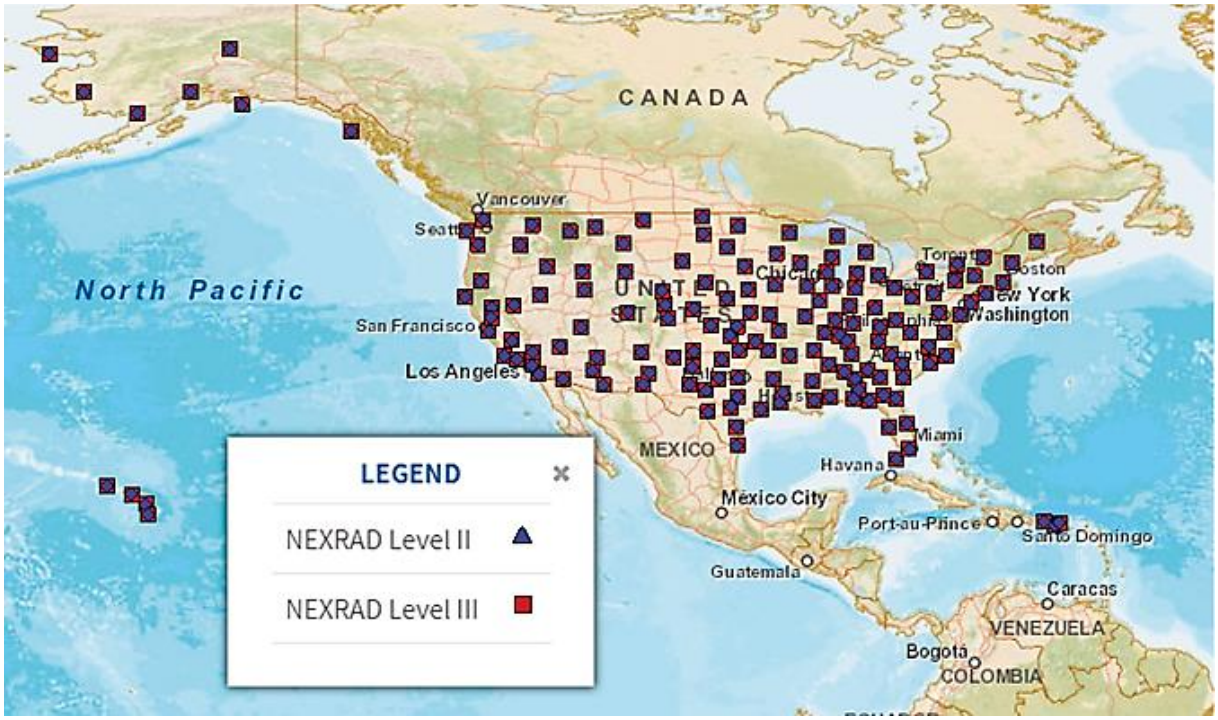
11.6. NEXRAD DAN RDA ASR

Berdasarkan sistem katalog yang diterapkan, kami melakukan beberapa percobaan untuk memvalidasi kelayakan pendekatan yang diusulkan. Kumpulan data untuk ilmu iklim umumnya berukuran sangat besar karena jangka waktunya panjang dan resolusi temporalnya tinggi. Kami mengambil dataset UCAR NEXRAD dan dataset RDA ASR (53,09 terabyte) sebagai contoh demonstrasi kami. Kemampuan pencarian pada dua kumpulan data dibuat di EarthCube CyberConnector. Kami melakukan serangkaian tes lengkap pada pencari dan hasilnya diperkenalkan di bawah.

Mencari Kumpulan Data NEXRAD

NEXRAD adalah kumpulan data yang sangat penting untuk penelitian ilmu iklim. Saat ini terdiri dari 160 lokasi di seluruh Amerika Serikat dan lokasi terpilih di luar negeri (seperti yang ditunjukkan pada Gambar 11.14). Kumpulan data dasar asli, termasuk tiga besaran data dasar meteorologi: Reflektivitas, kecepatan radial rata-rata, dan lebar spektrum, disebut Tingkat II. Produk turunannya disebut Level III, yang mencakup berbagai produk analisis meteorologi. Semua data NEXRAD Level-II tersedia melalui NCEI, serta penyedia cloud paket data besar NOAA, layanan web Amazon (http://thredds-aws.unidata.ucar.edu/thredds/catalog.html) dan Google Cloud (https://cloud.google.com/penyimpanan/dokumen/kumpulan data publik/nexrad). UCAR menyediakan data observasi hampir real-time melalui server data THREDDS mereka

(http://thredds.ucar.edu). Sayangnya, semua penyimpanan data ini masih belum dapat dicari saat ini, karena merupakan tantangan besar bagi katalog mana pun untuk mengindeks dan mencari file metadata dalam jumlah besar untuk catatan data radar yang sering diperbarui (setiap 6 menit). Kami menggunakan kumpulan data ini untuk membuktikan bahwa pendekatan katalogisasi yang diusulkan dapat bekerja dengan baik pada kumpulan data besar yang sering diperbarui.



Gambar 11.14. Peta Data Radar NCDC NOAA (NEXRAD Level II dan III).

Sistem yang lengkap terdiri dari layanan pemanen/pengindeks dan klien pencarian yang tersedia bagi pengguna sebagai aplikasi web. Hasilnya, pengguna dapat mencari beragam kumpulan data pengamatan dan pemodelan sistem Bumi yang heterogen secara bersamaan. Setelah metadata ditemukan, pengguna dapat menggunakan sistem visualisasi CyberConnector untuk secara bersamaan memvisualisasikan radar NEXRAD, pengamatan satelit, dan data produk model simulasi perkiraan hampir real-time. Karakteristik kinerja sistem dari pendekatan ini meningkat secara signifikan dibandingkan metode naif yang ada dalam mengumpulkan semua metadata kumpulan data.

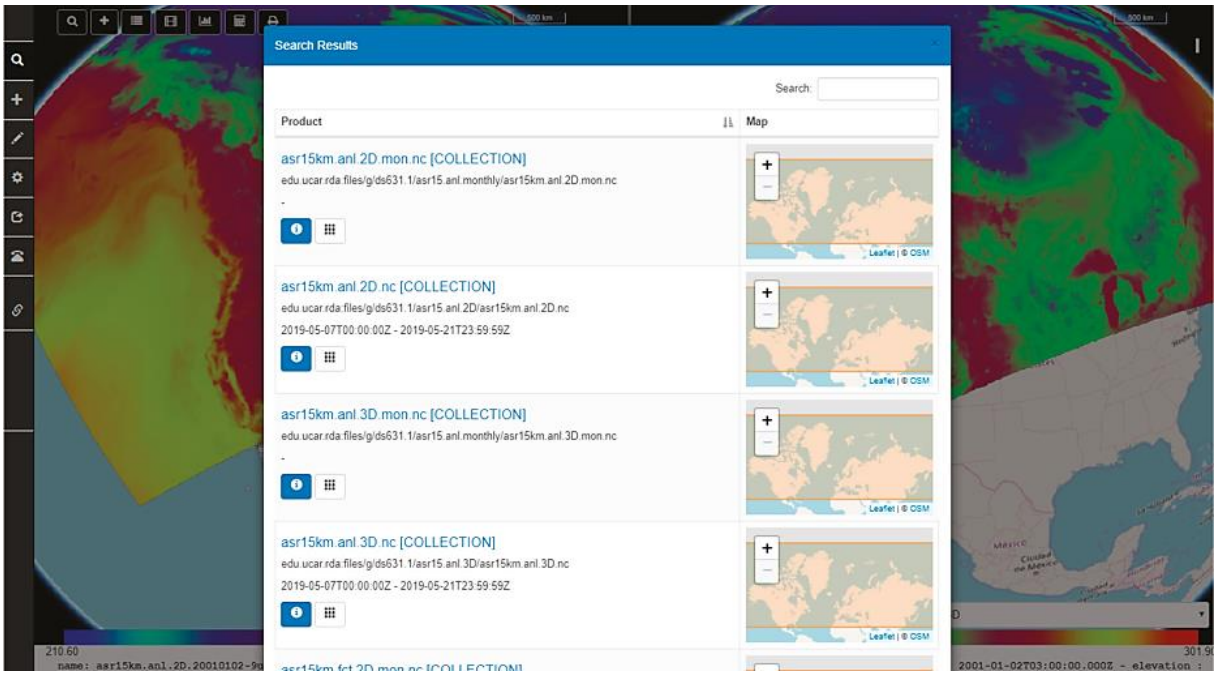
Mencari Repositori TDS UCAR RDA (Arsip Data Penelitian).

NCAR CISL (Computational & Information System Lab) yang didanai NSF mengelola Arsip Data Penelitian (RDA), yang menyimpan lebih dari 11.000 terabyte kumpulan data iklim dalam sistem penyimpanan data berkinerja tinggi.

RDA menampung banyak kumpulan data iklim saat ini, dan Analisis Ulang Sistem Arktik (ASR) adalah salah satunya. ASR adalah demonstrasi analisis ulang regional untuk wilayah Arktik yang lebih luas yang dikembangkan oleh Ohio State University. Dataset ASR versi 2 (versi terbaru) disajikan melalui RDA dengan total volume 53,04 terabyte. Resolusi horizontal adalah

15 km dan cakupan temporal dari tahun 2000 hingga 2016. Ia memiliki 34 tingkat tekanan (71 tingkat model), 31 permukaan (termasuk 3 variabel tanah), dan 11 variabel analisis udara atas, 71 permukaan (termasuk 3 variabel tanah)), dan 17 variabel prakiraan udara atas.

RDA menyediakan TDS untuk sebagian besar kumpulan data yang diarsipkannya. Kami mengambil metadata ASR dari TDS-nya dan menyediakannya untuk umum di CyberConnector. Seperti yang ditunjukkan pada Gambar 15, ilmuwan dapat mencari dataset ASR dengan memberikan kata kunci, luasan spasial, atau rentang waktu. Data ASR dalam format NetCDF yang dapat ditampilkan di COVALI. Kami mendemonstrasikan pencarian kumpulan data ASR di COVALI dan memvisualisasikan suhu pada 2 m di atas permukaan dalam waktu 12 jam. COVALI dan RDA dikerahkan di dua fasilitas yang didistribusikan dari jarak jauh. Interaksi antara penyimpanan data besar COVALI dan RDA dilakukan melalui antarmuka layanan standar dan melalui jaringan. Eksperimen tersebut membuktikan bahwa solusi yang diusulkan bekerja dengan baik untuk memungkinkan pencarian pada data besar jarak jauh.

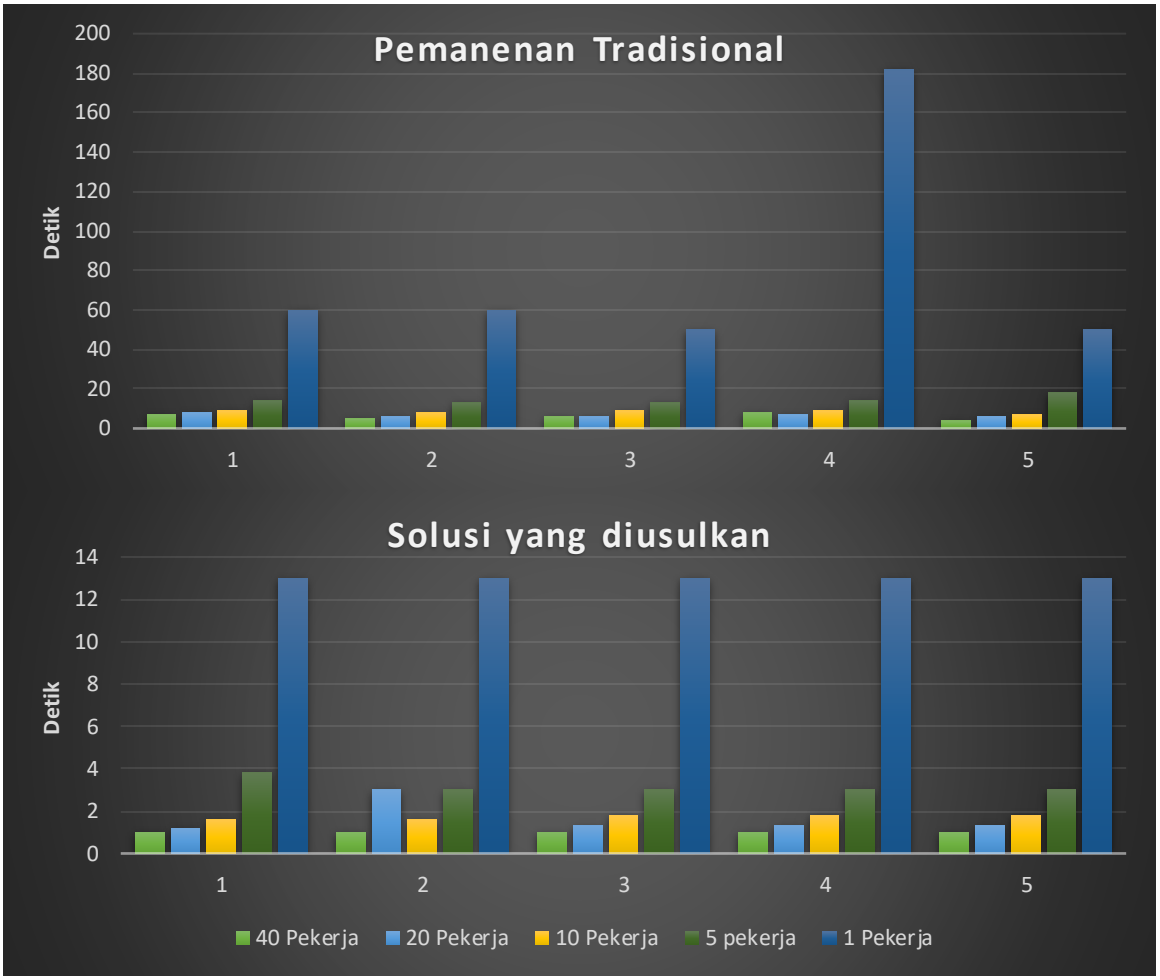


Gambar 11.15. Hasil penelusuran ASR (Arctic System Relysis) dan visualisasi suhu pada ketinggian 2 m di atas permukaan pada sistem visualisasi CyberConnector COVALI.

Evaluasi kinerja

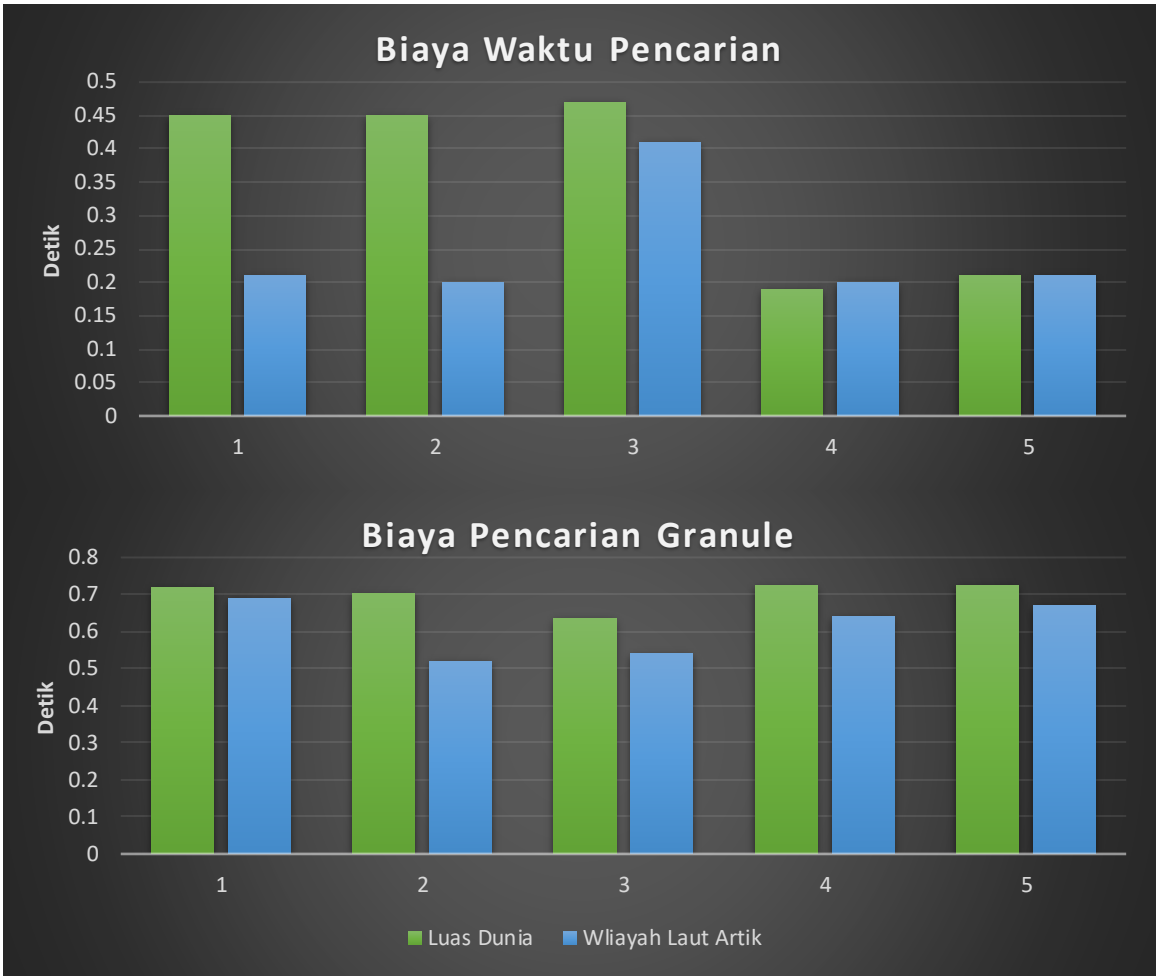
Pendekatan tradisional untuk membuat katalog kumpulan data iklim adalah dengan mengumpulkan seluruh file metadata dari setiap catatan data. Kami mengimplementasikan pencarian menggunakan metode tradisional sebelumnya, namun kinerjanya sangat lambat dan pengoperasian yang berkelanjutan tidak mungkin dilakukan untuk skenario praktis pembuatan katalog data besar. Setelah kami menerapkan strategi katalogisasi baru, kami mengujinya dengan merayapi beberapa ratus dan ribuan catatan dari UCAR THREDDS Data Server. Kami menguji menggunakan kelompok pekerja paralel yang berbeda: masing-masing 40 pekerja, 20 pekerja, 10 pekerja, 5 pekerja, dan satu pekerja, untuk mengukur peningkatan

perayapan paralel. Gambar 11.16 menampilkan biaya waktu pengujian untuk membandingkan kinerja pendekatan tradisional dan pendekatan yang diusulkan. Hasilnya menunjukkan bahwa pendekatan yang diusulkan mengungguli pendekatan tradisional setidaknya sepuluh kali lipat dari keseluruhan biaya waktu (dari ~10 hingga ~1 detik) dan memiliki peningkatan yang signifikan pada kecepatan pemanenan, penggunaan penyimpanan, dan kecepatan pencarian berdasarkan jumlah kumpulan data yang sedang diproses. diproses.



Gambar 11.16. Perbandingan kinerja (waktu dalam detik) dari pendekatan pemanenan tradisional (a) dan pendekatan kami (b), diambil sampelnya sebanyak 5 kali untuk merayapi 125 catatan.

Biaya waktu pencarian memiliki dua komponen. Waktu untuk mencari koleksi di katalog dan waktu untuk mengambil daftar butiran dari indeks butiran. Gambar 11.17 menunjukkan bahwa pengambilan hasil pencarian sangat cepat di sistem kami.



Gambar 11.17. Kinerja pencarian (waktu dalam detik) dengan dua parameter luasan spasial yang berbeda. (a) Waktu untuk menanyakan katalog (pencarian koleksi, langkah 1); (b) waktu untuk menanyakan indeks granul (pencarian granul, langkah 2).

Pencarian saat ini mendukung filter, termasuk kata kunci, format data, dan luasan spatiotemporal. Semuanya merupakan filter tetap dengan ketidakpastian yang lebih sedikit. Oleh karena itu, hasil yang dikembalikan tetap sama selama basis metadata tidak menambahkan data baru atau menghapus data yang sudah ada. Kelengkapan hasil 100% akurat karena pencatatan yang benar sesuai dengan kondisi filter. Pengguna dapat mempersempit cakupan spatiotemporal berdasarkan minat mereka dan memberikan satu atau lebih kata kunci yang dapat cocok dengan nama kolom data. Relevansi halaman pertama hasil pencarian bergantung pada hubungan antara kata kunci yang dimasukkan dan nilai kolom metadata. Berdasarkan pengalaman kami dengan para ilmuwan iklim, kami menemukan bahwa mereka biasanya tidak memasukkan kata kunci apa pun dan hanya menggunakan filter spatiotemporal untuk memeriksa apa yang dapat dicari dalam katalog. Begitu mereka memiliki wilayah yang diminati atau rentang waktu, mereka kemudian mendapat gambaran tentang kemungkinan data yang tersedia. Mereka membuka katalog hanya untuk menemukan URL akses untuk mengunduh atau memvisualisasikan file data. Biasanya, hasil dari klien penelusuran kami banyak karena filter longgar yang diberikan para ilmuwan. Hasil pada halaman pertama biasanya sangat berkaitan dengan kebutuhan para

ilmuwan. Pencarian yang lebih cerdas, seperti pencarian berbasis semantik, yang dapat menemukan hasil halaman pertama yang lebih akurat dengan relevansi yang lebih tinggi, akan dipelajari pada tahap pekerjaan berikutnya.

11.7 HAMBATAN IMPLEMENTASI

Solusi kami terhadap tantangan volume, variasi, dan kecepatan data besar yang dibahas dalam penelitian ini terdiri dari model metadata baru, serta arsitektur dan implementasi infrastruktur siber yang diturunkan dari model tersebut. Model metadata menggabungkan deskripsi konten metadata ("model informasi") dengan deskripsi struktur dan perilaku penyimpanan metadata. Infrastruktur siber terdiri dari layanan perayap yang memanfaatkan model metadata untuk mengoptimalkan strategi perayapan THREDDS guna menghilangkan transfer dan pemrosesan informasi metadata yang berlebihan. Selain itu, model repositori metadata memungkinkan layanan crawler melakukan transfer metadata tambahan, sehingga memungkinkan kemampuan pencarian waktu nyata. Infrastruktur siber yang ditunjukkan juga mencakup layanan katalog yang dapat dioperasikan yang menggunakan model metadata untuk meminimalkan penyimpanan informasi yang berlebihan. Terakhir, klien pencarian yang menggunakan katalog dan layanan crawler diimplementasikan.

Dapatkah Solusi yang Diusulkan Mengatasi Tantangan Volume?

Volume metadata ~25 GB untuk kumpulan data UCAR RADAR. Metode tradisional untuk mengumpulkan metadata mampu memproses sekitar satu catatan (dengan ukuran perkiraan 100 KB) per detik. Untuk sepenuhnya menyerap semua metadata THREDDS RADAR pada tingkat pemanenan yang diamati, diperlukan waktu 250.000 detik atau ~70 jam. Dengan menggunakan model metadata dan sistem katalogisasi yang diusulkan, kami mengamati tingkat pemanenan yang setidaknya 10 kali lebih cepat. Ini memungkinkan sinkronisasi harian semua metadata TDS Unidata.

Dapatkah Solusi yang Diusulkan Mengatasi Tantangan Kecepatan?

Kami menetapkan bahwa metadata RADAR baru (langsung) dihasilkan dengan kecepatan 330 catatan per menit. Kapasitas panen maksimum kami (dibatasi oleh kapasitas jaringan Unidata THREDDS) adalah 60 catatan per menit. Dengan menggunakan metode tradisional, kami tidak dapat mengimbangi kecepatan data. Dengan menggunakan pendekatan pengindeksan pemanen, kami dapat memproses hingga 1400 catatan per menit. Ini melebihi kecepatan produksi data THREDDS. Selain itu, dengan menggunakan pembaruan indeks tambahan selama pertukaran permintaan pencarian klien, kami dapat menargetkan proses pengumpulan pengindeksan ke sub-katalog yang tepat yang berisi informasi terbaru dan dengan demikian memberikan kemampuan pencarian real-time untuk data berkecepatan tinggi ini.

Dapatkah Solusi yang Diusulkan Mengurangi Redundansi Perayapan Metadata?

Solusi yang ditunjukkan di sini mampu mengurangi redundansi dalam perayapan dan konsumsi sumber daya penyimpanan. Misalnya, menggunakan metode tradisional dengan katalog Model Prakiraan, ~7000 catatan diunduh. Total penyimpanan yang digunakan adalah 1,85 GB. Metadata yang sama dapat diproses menggunakan pendekatan kami dengan

mengunduh hanya 45 contoh catatan metadata (2,2 MB) yang mewakili informasi tingkat pengumpulan. Ini mewakili pengurangan 99% dalam biaya transmisi dan penyimpanan data.

Apa Keuntungan dan Kerugian dari Solusi yang Diusulkan Dibandingkan dengan Strategi Pencarian Big Data Lainnya?

Solusinya menunjukkan manfaat yang diharapkan yang dijelaskan di awal penelitian ini. Kelemahan utama dari solusi ini adalah kompleksitas model dan sistem perangkat lunak. Perangkat lunak khusus harus dikembangkan untuk memproses katalog secara cerdas saat katalog tersebut dipanen. Untuk mendapatkan hasil yang lengkap dan akurat, metadata yang diserap harus dibersihkan dan diubah untuk mengisi informasi yang hilang dan membuatnya sesuai dengan model kami. Meskipun pendekatan kami cukup umum untuk bekerja dengan beberapa repositori TDS, dalam praktiknya, ketidakkonsistenan dan variasi tambahan dari setiap repositori harus direkonsiliasi menggunakan kode khusus. Pekerjaan kami menunjukkan bahwa membangun sistem katalog yang terpadu dan sangat efisien dapat dicari untuk repositori data sistem Bumi yang besar dan heterogen yang mendukung kueri real-time; namun, setiap solusi mempunyai keterbatasan dan biaya. Dalam hal ini, biayanya adalah kompleksitas dalam arsitektur perangkat lunak dan sistem, yang berarti peningkatan biaya pengembangan dan pemeliharaan perangkat lunak.

11.8 RINGKASAN

Pada bab ini mengusulkan dan mendemonstrasikan solusi katalogisasi baru berbasis infrastruktur siber untuk memungkinkan pencarian dua langkah yang efisien pada kumpulan data iklim besar dengan memanfaatkan pusat data yang ada dan teknologi layanan web tercanggih. Kami menggunakan kumpulan data besar yang disajikan oleh UCAR THREDDS Data Server (TDS), yang menyajikan data ESOM tingkat Petabyte dan memperbarui ratusan terabyte data setiap hari, sebagai kumpulan data studi kami untuk memvalidasi kelayakannya. Kami menganalisis struktur metadata di TDS dan membuat indeks untuk parameter data. Model registrasi metadata yang dikembangkan, yang mendefinisikan informasi konstan, membatasi informasi variabel, dan memanfaatkan koherensi spasial dan temporal dalam metadata, telah dibangun. Model ini memperoleh strategi pengambilan sampel untuk bot perayap web serentak berkinerja tinggi yang digunakan untuk mencerminkan metadata penting dari arsip data besar tanpa membebani sumber daya jaringan dan komputasi. Model metadata, crawler, dan layanan katalog yang sesuai standar membentuk infrastruktur siber pencarian tambahan, yang memungkinkan para ilmuwan melakukan pencarian hampir secara real-time dalam kumpulan data iklim besar. Kami bereksperimen dengan pendekatan pada UCAR TDS dan NCAR RDA TDS, dan hasilnya membuktikan bahwa pendekatan yang diusulkan mencapai tujuan desainnya, yang merupakan terobosan signifikan untuk server data iklim yang paling tidak dapat ditelusuri saat ini. Solusi ini mengidentifikasi informasi yang berlebihan dan menentukan frekuensi pengambilan sampel untuk menjaga bagian katalog sumber yang tidak dapat diprediksi tetap sinkron dengan katalog cermin hilir kami. Perayap-pengindeks hierarki otomatis dan sistem pencarian gratis menggunakan EarthCube CyberConnector yang sudah ada telah diterapkan. Perayapan metadata dan kinerja akses memvalidasi pendekatan

terintegrasi kami sebagai metode yang efektif untuk menghadapi tantangan data besar yang ditimbulkan oleh data Model dan Pengamatan Sistem Bumi yang heterogen dan real-time. Namun, meskipun pendekatan yang diusulkan lebih unggul dari solusi pencarian tradisional untuk data besar, pendekatan ini masih memakan waktu baik dalam proses perayapan maupun pencarian, dan mungkin ketinggalan zaman dalam menangani data streaming real-time. Di masa depan, kami akan mempelajari cara mengurangi waktu yang dihabiskan untuk meng-crawl metadata yang berlebihan dan menemukan metode berkinerja tinggi untuk penelusuran yang cepat dan cerdas.

DAFTAR PUSTAKA

Aachen, R.; Informatik, L.T.; Horrocks, I.; Sattler, U.; Tobies, S. Pspace-Algorithm for Deciding Alcnir+-Satisfiability. In LTCS-Report 98-08; ACM Digital Library: Aachen, Germany, 1998.

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv 2016, arXiv:1603.04467.

Abernathey, R.; Paul, K.; Hamman, J.; Rocklin, M.; Lepore, C.; Tippett, M.; Henderson, N.; Seager, R.; May, R.; Del Vento, D. Pangeo NSF Earthcube Proposal. Available online: https://figshare.com/articles/Pangeo_NSF_ Earthcube_Proposal/5361094 (accessed on 10 March 2020).

Agamennoni, G.; Nieto, J.I.; Nebot, E.M. Robust inference of principal road paths for intelligent transportation systems. IEEE Trans. Intell. Transp. Syst. 2011, 12, 298–308. [CrossRef]

Ahmed, M.; Fasy, B.T.; Hickmann, K.S.; Wenk, C. A path-based distance for street map comparison. ACM Trans. Spat. Algorithms Syst. 2015, 1, 1–28. [CrossRef]

Al-Areqi, S.; Lamprecht, A.-L.; Margaria, T. Automatic workflow composition in the geospatial domain: An application on sea-level rise impacts analysis. In Proceedings of the 19th AGILE International Conference on Geographic Information Science, Helsinki, Finland, 14–17 June 2016.

Al-Areqi, S.; Lamprecht, A.L.; Margaria, T. Constraints-driven automatic geospatial service composition: Workflows for the analysis of sea-level rise impacts. In Proceedings of the International Conference on Computational Science and Its Applications, Beijing, China, 4–7 July 2016; pp. 134–150.

Albrecht, J. Universal elementary GIS tasks-beyond low-level commands. In Proceedings of the Sixth International Symposium on Spatial Data Handling, Edinburgh, UK, 3–7 August 1994; pp. 209–222.

Alexander, C.; Deák, B.; Heilmeier, H. Micro-topography driven vegetation patterns in open mosaic landscapes. Ecol. Indic. 2016, 60, 906–920. [CrossRef]

Almeer, M.H. Cloud Hadoop Map Reduce For Remote Sensing Image Analysis. J. Emerg. Trends Comput. Inf. Sci. 2012, 3, 637–644.

Alsharif, A.A.A.; Pradhan, B. Urban Sprawl Analysis of Tripoli Metropolitan City (Libya) Using Remote Sensing Data and Multivariate Logistic Regression Model. J. Indian Soc. Remote Sens. 2014, 42, 149–163. [CrossRef]

Altintas, I.; Berkley, C.; Jaeger, E.; Jones, M. Kepler: An extensible system for design and execution of scientific workflows. In Proceedings of the International Conference on

Scientific and Statistical Database Management, Santorini Island, Greece, 23 June 2004; pp. 423–424.

Amatulli, G.; Domisch, S.; Tuanmu, M.-N.; Parmentier, B.; Ranipeta, A.; Malczyk, J.; Jetz, W. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Sci. Data 2018, 5, 180040. [CrossRef]

Ansari, S.; Del Greco, S.; Kearns, E.; Brown, O.; Wilkins, S.; Ramamurthy, M.; Weber, J.; May, R.; Sundwall, J.; Layton, J.; et al. Unlocking the Potential of NEXRAD Data through NOAA's Big Data Partnership. Bull. Am. Meteorol. Soc. 2018, 99, 189–204. [CrossRef]

Apache Jena. Available online: http://jena.apache.org/ (accessed on 29 August 2018).

Araya, Y.H.; Cabral, P. Analysis and modeling of urban land cover change in Setúbal and Sesimbra, Portugal. Remote Sens. 2010, 2, 1549–1563. [CrossRef]

Aronson, E.; Ferrini, V.; Gomez, B. Geoscience 2020: Cyberinfrastructure to Reveal the Past, Comprehend the Present, and Envision the Future; National Science Foundation: Alexandria, VA, USA, 2015.

Arul, U.; Prakash, S. A unified algorithm to automatic semantic composition using multilevel workflow orchestration. In Cluster Computing; Springer: New York, NY, USA, 2018; pp. 1–22.

Arvor, D.; Durieux, L.; Andrés, S.; Laporte, M.-A. Advances in Geographic Object-Based Image Analysis with ontologies: A review of main contributions and limitations from a remote sensing perspective. ISPRS J. Photogramm. Remote. Sens. 2013, 82, 125–137. [CrossRef]

Baader, F.; Sattler, U. An Overview of Tableau Algorithms for Description Logics. Stud. Log. 2001, 69, 5–40. [CrossRef]

Bacani, V.M.; Sakamoto, A.Y.; Quénol, H.; Vannier, C.; Corgne, S. Markov chains-cellular automata modeling and multicriteria analysis of land cover change in the Lower Nhecolândia subregion of the Brazilian Pantanal wetland. J. Appl. Remote Sens. 2016, 10, 016004. [CrossRef]

Bai, Y.; Di, L. Providing access to satellite imagery through OGC catalog service interfaces in support of the Global Earth Observation System of Systems. Comput. Geosci. 2011, 37, 435–443. [CrossRef]

Bai, Y.; Di, L.; Chen, A.; Liu, Y.; Wei, Y. Towards a Geospatial Catalogue Federation Service. Photogramm. Eng. Remote Sens. 2007, 73, 699–708. [CrossRef]

Bai, Y.; Di, L.; Wei, Y. A taxonomy of geospatial services for global service discovery and interoperability. Comput. Geosci. 2009, 35, 783–790. [CrossRef]

Bala, M.; Boussaid, O.; Alimazighi, Z. A Fine Grained Distribution Approach for ETL Processes in Big data Environments. Data Knowl. Eng. 2017, 111, 114–136. [CrossRef]

Bala, M.; Boussaid, O.; Alimazighi, Z. P-ETL: Parallel-ETL based on the MapReduce paradigm. In Proceedings of the 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), Doha, Qatar, 10–13 November 2014; pp. 42–49.

Batty, M.; Xie, Y.; Sun, Z. Modeling urban dynamics through GIS-based cellular automata. Comput. Environ. Urban Syst. 1999, 23, 205–233. [CrossRef]

Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A neural probabilistic language model. J. Mach. Learn. Res. 2003, 3, 1137–1155.

Bensmann, F.; Alcacerlabrador, D.; Ziegenhagen, D.; Roosmann, R. The richwps environment for orchestration. ISPRS Int. J. Geo-Inf. 2014, 3, 1334–1351. [CrossRef]

Bernard, L.; Mäs, S.; Müller, M.; Henzen, C.; Brauner, J. Scientific geodata infrastructures: Challenges, approaches and directions. Int. J. Digit. Earth 2014, 7, 613–633. [CrossRef]

Bhattacharya, A.; Culler, D.E.; Ortiz, J.; Hong, D.; Whitehouse, K.; Culler, D. Enabling Portable Building Applications through Automated Metadata Transformation; University of California at Berkeley: Berkeley, CA, USA, 2014.

Biagioni, J.; Eriksson, J. Inferring road maps from global positioning system traces. Transp. Res. Rec. J. Transp. Res. Board 2015, 2291, 61–71. [CrossRef]

Biagioni, J.; Eriksson, J. Map inference in the face of noise and disparity. In Proceedings of the International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 6–9 November 2012.

Bindzárová Gergel'ová, M.; Kuzevicˇová, Ž.; Labant, S.; Gašinec, J.; Kuzevicˇ, Š.; Unucka, J.; Liptai, P. Evaluation of Selected Sub-Elements of Spatial Data Quality on 3D Flood Event Modeling: Case Study of Prešov City, Slovakia. Appl. Sci. 2020, 10, 820. [CrossRef]

Bogaart, P.W.; Troch, P.A. Curvature distribution within hillslopes and catchments and its effect on the hydrological response. Hydrol. Earth Syst. Sci. 2006, 10, 925–936. [CrossRef]

Bollacker, K.; Cook, R.; Tufts, P. Freebase: A Shared Database of Structured General Human Knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22–26 July 2007.

Boucher, C.; Noyer, J.C. Automatic detection of topological changes for digital road map updating. IEEE Trans. Instrum. Meas. 2012, 61, 3094–3102. [CrossRef]

Brainerd, J.; Pang, A. Interactive map projections and distortion. Comput. Geosci. 2001, 27, 299–314. [CrossRef]

Brauner, J. Formalizations for Geooperators-Geoprocessing in Spatial Data Infrastructures. Ph.D. Thesis, Technische Universität Dresden, Dresden, Germany, 2015.

Bresenham, J.E. Algorithm for computer control of a digital plotter. IBM Syst. J. 1999, 4, 25–30. [CrossRef]

Brinkhoff, T.; Kriegel, H.-P.; Schneider, R.; Braun, A. Measuring the Complexity of Polygonal Objects. In Proceedings of the ACM-GIS, Baltimore, MD, USA, 1–2 December 1995; p. 109.

Bromwich, D.H.; Wilson, A.B.; Bai, L.; Liu, Z.; Barlage, M.; Shih, C.-F.; Maldonado, S.; Hines, K.M.; Wang, S.-H.; Woollen, J.; et al. The Arctic System Reanalysis, Version 2. Bull. Am. Meteorol. Soc. 2018, 99, 805–828. [CrossRef]

Brown, S.H. Knowledge Representation and the Logical Basis of Ontology; Springer: London, UK, 2012; pp. 11–50.

Bryson, N.; Mobolurin, A. Towards modeling the query processing relevant shape complexity of 2D polygonal spatial objects. Inf. Softw. Technol. 2000, 42, 357–365. [CrossRef]

Bu, R.C.; Chang, Y.; Hu, Y.M.; Li, X.Z.; He, H.S. Measuring spatial information changes using Kappa coefficients: A case study of the city groups in central Liaoning province. Acta Ecol. Sin. 2005, 205, 4.

Burnett, K.; Ng, K.B.; Park, S. A comparison of the two traditions of metadata development. J. Am. Soc. Inf. Sci. 1999, 50, 1209–1217. [CrossRef]

Calvin, K.; Bond-Lamberty, B. Integrated human-earth system modeling—State of the science and future directions. Environ. Res. Lett. 2018, 13, 063006. [CrossRef]

Camras, L.A. An Event—Emotion or Event—Expression Hypothesis? A Comment on the Commentaries on Bennett, Bendersky, and Lewis (2002). Infancy 2010, 6, 431–433. [CrossRef]

Cao, L.; Krumm, J. From GPS traces to a routable road map. In Proceedings of the Workshop on Advances in Geographic Information Systems, New York, NY, USA, 4–6 November 2009; pp. 3–12.

CARVER, S.J. Integrating multi-criteria evaluation with geographical information systems. Int. J. Geogr. Inf. Syst. 1991, 5, 321–339. [CrossRef]

Chae, J.; Thom, D.; Jang, Y.; Kim, S.; Ertl, T.; Ebert, D.S. Special Section on Visual Analytics: Public behavior response analysis in disaster events utilizing visual analytics of microblog data. Comput. Graph. 2014, 38, 51–60. [CrossRef]

Chen, C.; Cheng, Y. Roads Digital Map Generation with Multi-track GPS Data. IEEE Comput. Soc. 2008, 12. [CrossRef]

Chen, J.; Deng, S.; Chen, H. Crowdgeokg: Crowdsourced Geo-Knowledge Graph. In Proceedings of the China Conference on Knowledge Graph and Semantic Computing, Chengdu, China, 26–29 August 2017.

Chen, N.; Di, L.; Yu, G.; Gong, J.; Wei, Y. Use of ebRIM-based CSW with sensor observation services for registry and discovery of remote-sensing observations. Comput. Geosci. 2009, 35, 360–372. [CrossRef]

Chen, P.; Shi, W. Measuring the Spatial Relationship Information of Multi-Layered Vector Data. ISPRS Int. J. Geo-Inf. 2018, 7, 88. [CrossRef]

Chen, Y.; Krumm, J. Probabilistic modeling of traffic lanes from GPS traces. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 81–88.

Chen, Y.; Sundaram, H. Estimating complexity of 2D shapes. In Proceedings of the 2005 IEEE 7th Workshop on Multimedia Signal Processing, Shanghai, China, 30 October–2 November 2005; pp. 1–4.

Chen, Y.; Zhou, L.; Tang, Y.; Singh, J.P.; Bouguila, N.; Wang, C.; Wang, H.; Du, J. Fast neighbor search by using revised kd tree. Inf. Sci. 2019, 472, 145–162. [CrossRef]

Chen, Z.; Chen, N. Use of service middleware based on ECHO with CSW for discovery and registry of MODIS data. Geo-Spat. Inf. Sci. 2010, 13, 191–200. [CrossRef]

Chen, Z.; Ma, L.; Liang, W. Polygon Overlay Analysis Algorithm Based on Monotone Chain and STR Tree in the Simple Feature Model. In Proceedings of the 2010 International Conference on Electrical & Control Engineering, Wuhan, China, 25–27 June 2010.

Cheng, Y. Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell. 1995, 17, 790–799. [CrossRef]

Clarke, K.C.; Hoppen, S.; Gaydos, L. A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. Environ. Plan. B Plan. Des. 1997, 24, 247–261. [CrossRef]

Cohen, J.; Dolan, B.; Dunlap, M.; Hellerstein, J.M.; Welton, C. MAD skills: New analysis practices for big data. Proc. VLDB Endow. 2009, 2, 1481–1492. [CrossRef]

Collobert, R.; Weston, J. A unified architecture for natural language processing:deep neural networks with multitask learning. In Proceedings of the International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 160–167.

Costa, G.H.R.; Baldo, F. Generation of road maps from trajectories collected with smartphone—A method based on genetic algorithm. Appl. Soft Comput. 2015, 37, 799–808. [CrossRef]

Couclelis, H. Ontologies of geographic information. Int. J. Geogr. Inf. Sci. 2010, 24, 1785–1809. [CrossRef]

Coyle, D.; Meier, P. New technologies in emergencies and conflicts: The role of information and social networks. Washington D 2009. Available online: https://www.popline.org/node/209135 (accessed on 12 January 2019).

Crowley, M.A.; Cardille, J.A.; White, J.C.; Wulder, M.A. Multi-sensor, multi-scale, Bayesian data synthesis for mapping within-year wildfire progression. Remote Sens. Lett. 2019, 10, 302–311. [CrossRef]

Crubézy, M.; Musen, M.A. Ontologies in Support of Problem Solving; Springer: Berlin/Heidelberg, Germany, 2004; pp. 321–341.

CyberinfrastruCture Vision for 21st Century DisCoVery; National Science Foundation Cyberinfrastructure Council: Arlington, VA, USA, 2007.

Dang, L.M.; Ibrahim Hassan, S.; Suhyeon, I.; Sangaiah, A.K.; Mehmood, I.; Rho, S.; Seo, S.; Moon, H. UAV based wilt detection system via convolutional neural networks. Sustain. Comput. Inform. Syst. 2018, in press. [CrossRef]

Danielson, J.J.; Gesch, D.B. Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010); U.S. Geo-logical Survey Open-File Report 2011–1073; United States Geological Survey (USGS): Sioux Falls, SD, USA, 2011.

Davies, J.J.; Beresford, A.R.; Hopper, A. Scalable, distributed, real-time map generation. IEEE Pervasive Comput. 2006, 5, 47–54. [CrossRef]

Davis, R. What Is a Knowledge Representation? AI Mag. 1993, 14, 17–33.

Day, H. Evaluations of subjective complexity, pleasingness and interestingness for a series of random polygons varying in complexity. Percept. Psychophys. 1967, 2, 281–286. [CrossRef]

Dean, J.; Ghemawat, S. MapReduce. Commun. ACM 2008, 51, 107. [CrossRef]

Demattê, J.A.M.; Safanelli, J.L.; Poppiel, R.R.; Rizzo, R.; Silvero, N.E.Q.; Mendes, W.S.; Bonfatti, B.R.; Dotto, A.C.; Salazar, D.F.U.; Mello, F.A.O.; et al. Bare Earth's Surface Spectra as a Proxy for Soil Resource Monitoring. Sci. Rep. 2020, 10, 4461. [CrossRef]

Desai, K.; Devulapalli, V.; Agrawal Asst, S.; Kathiria Asst, P.; Patel Professor, A. Web Crawler: Review of Different Types of Web Crawler, Its Issues, Applications and Research Opportunities. Int. J. Adv. Res. Comput. Sci. 2017, 8, 1199–1202.

Devi, P.S.; Rao, V.V.; Raghavender, K. Emerging Technology Big data-Hadoop over Datawarehousing ETL. In Proceedings of the International Conference (IRF), Pretoria, South Africa, 2–4 September 2014; pp. 30–34.

Di, L. Distributed geospatial information services-architectures, standards, and research issues. Int Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 2004, 35 Pt 2.

Di, L. Geospatial Sensor Web and Self-adaptive Earth Predictive Systems (SEPS). In Proceedings of the Earth Science Technology Office (ESTO)/Advanced Information System Technology (AIST) Sensor Web Principal Investigator (PI) Meeting, San Diego, CA, USA, 13 February 2007.

Di, L. The development of remote-sensing related standards at FGDC, OGC, and ISO TC 211. In Proceedings of the 2003 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2003), Toulouse, France, 21–25 July 2003; Volume 1, pp. 643–647.

Di, L.; Kobler, B. NASA Standards for Earth Remote Sensing Data. Int. Arch. Photogramm. Remote Sens. 2000, 33, 147–155.

Di, L.; Moe, K.L.; Yu, G. Metadata requirements analysis for the emerging Sensor Web. Int. J. Digit. Earth 2009, 2, 3–17. [CrossRef]

Di, L.; Schlesinger, B.M.; Kobler, B.U.S. Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata; Federal Geographic Data Committee: Reston, VA, USA, 2000.

Di, L.; Shao, Y.; Kang, L. Implementation of Geospatial Data Provenance in a Web Service Workflow Environment with ISO 19115 and ISO 19115-2 Lineage Model. IEEE Trans. Geosci. Remote Sens. 2013, 51, 5082–5089.

Di, L.; Sun, Z.; Yu, E.; Song, J.; Tong, D.; Huang, H.; Wu, X.; Domenico, B. Coupling of Earth science models and earth observations through OGC interoperability specifications. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 3602–3605.

Di, L.; Sun, Z.; Zhang, C. Facilitating the Easy Use of Earth Observation Data in Earth system Models through CyberConnector. In Proceedings of the AGU Fall Meeting, New Orleans, LA, USA, 11–15 December 2017. Abstract #IN21D-0072.

Di, L.; Yu, G.; Shao, Y.; Bai, Y.; Deng, M.; McDonald, K.R. Persistent WCS and CSW services of GOES data for GEOSS. In Proceedings of the 2010 IEEE International Geoscience and Remote Sensing Symposium, Honolulu, HI, USA, 25–30 July 2010; pp. 1699–1702.

Dietzel, C.; Clarke, K.C. Toward optimal calibration of the SLEUTH land use change model. Trans. GIS 2007, 11, 29–45. [CrossRef]

Ding, Y.; Foo, S. Ontology research and development. Part 1: A review of ontology generation. J. Inf. Sci. 2002, 28, 123–136.

Domenico, B.; Caron, J.; Davis, E.; Kambic, R.; Nativi, S. Thematic Real-Time Environmental Distributed Data Services (THREDDS): Incorporating Interactive Analysis Tools into NSDL; Multimedia Research Group, University of Southampton: Southampton, UK, 1997; Volume 2.

Donchyts, G.; Baart, F.; Winsemius, H.; Gorelick, N.; Kwadijk, J.; van de Giesen, N. Earth's surface water change over the past 30 years. Nat. Clim. Chang. 2016, 6, 810–813. [CrossRef]

Du, Z.; Li, W.; Zhou, D.; Tian, L.; Ling, F.; Wang, H.; Gui, Y.; Sun, B. Analysis of Landsat-8 OLI imagery for land surface water mapping. Remote Sens. Lett. 2014, 5, 672–681. [CrossRef]

Edelkamp, S.; Schrödl, S. Route Planning and Map Inference with Global Positioning Traces. In Computer Science in Perspective, Essays Dedicated to Thomas Ottmann; Springer: Berlin, Heidelberg, 2003; pp. 128–151.

Ekman, P. Basic Emotions. In Handbook of Cognition & Emotion; Dalgleish, T., Power, M.J., Eds.; John Wiley & Sons: Hoboken, NY, USA, 1999.

El Akkaoui, Z.; Zimanyi, E. Defining ETL worfklows using BPMN and BPEL. In DOLAP '09, Proceedings of the ACM Twelfth International Workshop on Data Warehousing and OLAP, Hong Kong, China, 6 November 2009; ACM: New York, NY, USA, 2009; pp. 41–48.

Eldawy, A. SpatialHadoop: Towards flexible and scalable spatial processing using MapReduce. In Proceedings of the SIGMOD Ph.D. Symposium 2014, Snowbird, UT, USA, 22 June 2014; pp. 46–50.

Eldawy, A. SpatialHadoop: Towards flexible and scalable spatial processing using MapReduce. In Proceedings of the SIGMOD PhD symposium 2014, Snowbird, UT, USA, 22 June 2014; pp. 46–50.

Eldawy, A.; Alarabi, L.; Mokbel, M.F. Spatial partitioning techniques in SpatialHadoop. Proc. VLDB Endow. 2015, 8, 1602–1605. [CrossRef]

Eldawy, A.; Mokbel, M.F. A demonstration of spatialhadoop: An efficient mapreduce framework for spatial data. Proc. VLDB Endow. 2013, 6, 1230–1233. [CrossRef]

Eldawy, A.; Mokbel, M.F. Spatialhadoop: A mapreduce framework for spatial data. In Proceedings of the 2015 IEEE 31st International Conference on Data Engineering, Seoul, Korea, 13–17 April 2015; pp. 1352–1363.

El-Sheimy, N.; Schwarz, K.P. Navigating urban areas by VISAT—A mobile mapping system integrating GPS/INS/digital cameras for GIS applications. Navigation 1998, 45, 275–285. [CrossRef]

Endo, Y.; Toda, H.; Nishida, K.; Kawanobe, A. Deep feature extraction from trajectories for transportation mode estimation. In Pacific-Asia Conference on Knowledge Discovery and Data Mining; Springer: Berlin, Germany, 2016; pp. 54–66.

Etemad, M.; Soares Júnior, A.; Matwin, S. Predicting Transportation Modes of GPS Trajectories using Feature Engineering and Noise Removal. In Advances in Artificial Intelligence: 31st

Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8–11, 2018, Proceedings 31; Springer International Publishing: New York, NY, USA, 2018; pp. 259–264.

Faloutsos, C.; Kamel, I. Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In Proceedings of the Thirteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Minneapolis, MN, USA, 24–27 May 1994; pp. 4–13.

Farnaghi, M.; Mansourian, A. Disaster planning using automated composition of semantic OGC web services: A case study in sheltering. Comput. Environ. Urban Syst. 2013, 41, 204–218. [CrossRef]

Florinsky, I.V. An illustrated introduction to general geomorphometry. Prog. Phys. Geogr. Earth Env. 2017, 41, 723–752. [CrossRef]

Florinsky, I.V. Digital Terrain Analysis in Soil Science and Geology; Academic Press: Cambridge, MA, USA, 2016; ISBN 9780128046326.

Florinsky, I.V. Spheroidal equal angular DEMs: The specificity of morphometric treatment. Trans. GIS 2017, 21, 1115–1129. [CrossRef]

Florinsky, I.V.; Pankratov, A.N. A universal spectral analytical method for digital terrain modeling. Int. J. Geogr. Inf. Sci. 2016, 30, 2506–2528. [CrossRef]

Fonseca, F.T.; Egenhofer, M.J. Ontology-driven geographic information systems. In Proceedings of the 7th ACM International Symposium on Advances in Geographic Information Systems, Kansas City, MO, USA, 2–6 November 1999; Volume 71, pp. 14–19.

Foody, G.M. Map comparison in GIS. Prog. Phys. Geogr. 2007, 31, 439–445. [CrossRef]

Freitas, T.R.M.; Coelho, A.; Rossetti, R.J.F. Correcting routing information through GPS data processing. In Proceedings of the International IEEE Conference on Intelligent Transportation Systems, Piscataway, NJ, USA, 19–22 September 2010; pp. 706–711.

Freitas, T.R.M.; Coelho, A.; Rossetti, R.J.F. Improving digital maps through GPS data processing. In Proceedings of the International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO, USA, 12–15 October 2009; pp. 1–6.

Fu, X.; Wang, X.; Yang, Y.J. Deriving suitability factors for CA-Markov land use simulation model based on local historical data. J. Environ. Manag. 2018, 206, 10–19.

Gaigalas, J.; Di, L.; Sun, Z. Advanced Cyberinfrastructure to Enable Search of Big Climate Datasets in THREDDS. ISPRS Int. J. Geo-Inf. 2019, 8, 494. [CrossRef]

Gauss, C.F. General Investigations of Curved Surfaces of 1827 and 1825; Princeton University Library: Princeton, NY, USA, 1902.

Goldmann, E.; Galea, S. Mental health consequences of disasters. Ann. Rev. Public Health 2014, 35, 169. [CrossRef] [PubMed]

Golledge, R.G. The Nature of Geographic Knowledge. Ann. Assoc. Am. Geogr. 2015, 92, 1–14. [CrossRef]

Goltz, J.D.; Russell, L.A.; Bourque, L.B. Initial behavioural response to a rapid onset disaster. Int. J. Mass Emerg. Disasters 1992, 10, 43–69.

Goodchild, M.F. Citizens as sensors: Web 2.0 and the volunteering of geographic information. GeoFocus. Rev. Int. Ciencia Tecnol. Inf. Geogr. 2007, 7, 8–10.

Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Env. 2017, 202, 18–27. [CrossRef]

Gorton, S.; Reiff-Marganiec, S. Towards a task-oriented, policy-driven business requirements specification for web services. In Proceedings of the International Conference on Business Process Management, Vienna, Austria, 5–7 September 2006; pp. 465–470.

Greenberg, J. Metadata and the World Wide Web. Encycl. Libr. Inf. Sci. 2003, 3, 1876–1888.

Greiner, G.; Hormann, K. Efficient clipping of arbitrary polygons. ACM Trans. Graph. (TOG) 1998, 17, 71–83. [CrossRef]

Gruber, T.R. A translational approach to portable ontologies. Knowl. Acquis. 1993, 5, 199–220. [CrossRef]

Gruber, T.R. Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum.- Comput. Stud. 1995, 43, 907–928. [CrossRef]

Gruebner, O.; Lowe, S.R.; Sykora, M.; Shankardass, K.; Subramanian, S.V.; Galea, S. A novel surveillance approach for disaster mental health. PLoS ONE 2017, 12, e0181233. [CrossRef] [PubMed]

Guan, D.; Gao, W.; Watari, K.; Fukahori, H. Land use change of Kitakyushu based on landscape ecology and Markov model. J. Geogr. Sci. 2008, 18, 455–468. [CrossRef]

Guan, Q.; Clarke, K.C. A general-purpose parallel raster processing programming library test application using a geographic cellular automata model. Int. J. Geogr. Inf. Sci. 2010, 24, 695–722. [CrossRef]

Guan, Q.; Shi, X.; Huang, M.; Lai, C. A hybrid parallel cellular automata model for urban growth simulation over GPU/CPU heterogeneous architectures. Int. J. Geogr. Inf. Sci. 2016, 30, 494–514. [CrossRef]

Guan, Q.; Zeng, W.; Gong, J.; Yun, S. pRPL 2.0: Improving the parallel raster processing library. Trans. GIS 2014, 18, 25–52. [CrossRef]

Guarino, N.; Oberle, D.; Staab, S. What Is an Ontology? Handb. Ontol. 2009, 1–17. [CrossRef]

Guest, M. An overview of vector and parallel processors in scientific computation. J. Comput. Phys. Commun. 1989, 57, 560. [CrossRef]

Guo, T.; Iwamura, K.; Koga, M. Towards high accuracy road maps generation from massive GPS Traces data. In Proceedings of the 2007 IEEE International Geoscience and Remote Sensing Symposium, Barcelona, Spain, 23–28 July 2007; pp. 667–670.

Habermann, T. Metadata Life Cycles, Use Cases and Hierarchies. Geosciences 2018, 8, 179. [CrossRef]

Halmy, M.W.A.; Gessler, P.E.; Hicke, J.A.; Salem, B.B. Land use/land cover change detection and prediction in the north-western coastal desert of Egypt using Markov-CA. Appl. Geogr. 2015, 63, 101–112. [CrossRef]

Hansen, M.C.; Potapov, P.V.; Moore, R.; Hancher, M.; Turubanova, S.A.; Tyukavina, A.; Thau, D.; Stehman, S.V.; Goetz, S.J.; Loveland, T.R.; et al. High-Resolution Global Maps of 21st-Century Forest Cover Change. Science 2013, 342, 850–853. [CrossRef] [PubMed]

Haslhofer, B.; Klas, W. A survey of techniques for achieving metadata interoperability. ACM Comput. Surv. 2010, 42, 7. [CrossRef]

Haubrich, H. International Charter on Geographical Education. J. Geogr. 1997, 96, 33–39.

He, Z.; Liu, Q.; Deng, M.; Xu, F. Handling multiple testing in local statistics of spatial association by controlling the false discovery rate: A comparative analysis. In Proceedings of the IEEE 2nd International Conference 2017Big data Analysis (ICBDA), Beijing, China, 10–12 March 2017; pp. 684–687.

Heiss, W.H.; McGrew, D.L.; Sirmans, D. Nexrad: Next generation weather radar (WSR-88D). Microw. J. 1990, 33, 79–89.

Hishe, S.; Bewket, W.; Nyssen, J.; Lyimo, J. Analysing past land use land cover change and CA-Markov-based future modelling in the Middle Suluh Valley, Northern Ethiopia. Geocarto Int. 2019, 1–31. [CrossRef]

Hofer, B.; Brauner, J.; Jackson, M.; Granell, C.; Rodrigues, A.; Nüst, D.; Wiemann, S. Descriptions of spatial operations—Recent approaches and community feedback. Int. J. Spat. Data Infrastruct. Res. 2015, 10, 124–137.

Hofer, B.; Mãs, S.; Brauner, J.; Bernard, L. Towards a knowledge base to support geoprocessing workflow development. Int. J. Geogr. Inf. Syst. 2016, 31, 694–716. [CrossRef]

Hofer, B.; Papadakis, E.; Mäs, S. Coupling knowledge with GIS operations: The benefits of extended operation descriptions. Int. J. Geo-Inf. 2017, 6, 40. [CrossRef]

Hoffart, J.; Suchanek, F.M.; Berberich, K.; Weikum, G. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell. 2013, 194, 28–61. [CrossRef]

Holm, P.; Goodsite, M.E.; Cloetingh, S.; Agnoletti, M.; Moldan, B.; Lang, D.J.; Leemans, R.; Moeller, J.O.; Buendía, M.P.; Pohl, W.; et al. Collaboration between the natural, social and human sciences in Global Change Research. Environ. Sci. Policy 2013, 28, 25–35. [CrossRef]

Horrocks, I.; Sattler, U.; Tobies, S. Practical Reasoning for Expressive Description Logics. In Proceedings of the International Conference on Logic for Programming and Automated Reasoning, Tbilisi, Georgia, 6–10 September 1999.

Horrocks, I.; Sattler, U.; Tobies, S. Practical reasoning for very expressive description logics. Log. J. IGPL 2000, 8, 239–263. [CrossRef]

Hu, C.; Di, L.; Yang, W. The research of interoperability in spatial catalogue service between CSW and THREDDS. In Proceedings of the 2009 17th International Conference on Geoinformatics, Fairfax, VA, USA, 12–14 August 2009; pp. 1–5.

Hu, L.; Yue, P.; Zhang, M.; Gong, J.; Jiang, L.; Zhang, X. Task-oriented sensor web data processing for environmental monitoring. Earth Sci. Inform. 2015, 8, 511–525. [CrossRef]

Huang, C.-W.; Shih, T.-Y. On the complexity of point-in-polygon algorithms. Comput. Geosci. 1997, 23, 109–118. [CrossRef]

Hull, D.; Wolstencroft, K.; Stevens, R.; Goble, C.; Pocock, M.R.; Li, P.; Oinn, T. Taverna: A tool for building and running workflows of services. Nucleic Acids Res. 2006, 34, 729–732. [CrossRef] [PubMed]

Hurrell, J.W.; Holland, M.M.; Gent, P.R.; Ghan, S.; Kay, J.E.; Kushner, P.J.; Lamarque, J.-F.; Large, W.G.; Lawrence, D.; Lindsay, K.; et al. The Community Earth system Model: A Framework for Collaborative Research. Bull. Am. Meteorol. Soc. 2013, 94, 1339–1360. [CrossRef]

Int. J. U- E-Serv. Sci. Technol. 2012, 5, 43–58.

ISO. ISO 19115: Geographic Information—Metadata; ISO: Geneva, Switzerland, 2013.

ISO. ISO 19123: Geographic Information—Schema for Coverage Geometry and Functions; The International Organization for Standardization: Geneva, Switzerland, 2005.

J. Geogr. Inf. Syst. 2012, 4, 542–554. [CrossRef]

Jang, S.; Kim, T.; Lee, E. Map generation system with lightweight GPS trace data. In Proceedings of the International Conference on Advanced Communication Technology, Miyazaki, Japan, 23–25 June 2010; Volume 2, pp. 1489–1493.

Jantz, C.A.; Goetz, S.J.; Shelley, M.K. Using the SLEUTH urban growth model to simulate the impacts of future policy scenarios on urban land use in the Baltimore-Washington metropolitan area. Environ. Plan. B Plan. Des. 2004, 31, 251–271. [CrossRef]

JAXA EORC. ALOS Global Digital Surface Model "ALOS World 3D-30m (AW3D30)". Available online: https://www.eorc.jaxa.jp/ALOS/en/aw3d30/index.htm (accessed on 4 April 2020).

Jeansoulin, R. Review of forty years of technological changes in geomatics toward the big data paradigm. ISPRS Int. J. Geo-Inf. 2016, 5, 155. [CrossRef]

Jia, K.; Wei, X.; Gu, X.; Yao, Y.; Xie, X.; Li, B. Land cover classification using Landsat 8 Operational Land Imager data in Beijing, China. Geocarto Int. 2014, 29, 941–951. [CrossRef]

Jiang, L.; Sun, Z.; Qi, Q.; Zhang, A. Spatial Correlation between Traffic and Air Pollution in Beijing. Prof. Geogr. 2019, 71, 654–667. [CrossRef]

Jo, J.; Lee, K.W. High-Performance Geospatial Big data Processing System Based on MapReduce. ISPRS Int. J. Geo-Inf. 2018, 7, 399. [CrossRef]

Jo, J.; Lee, K.-W. Map Reduce-Based D_ELT Framework to Address the Challenges of Geospatial Big Data. ISPRS Int. J. Geo-Inf. 2019, 8, 475. [CrossRef]

John Caron, U.; Davis, E. UNIDATA's THREDDS data server. In Proceedings of the 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology, Atlanta, GA, USA, 27 January–3 February 2006.

Jozefowicz, R.; Vinyals, O.; Schuster, M.; Shazeer, N.; Wu, Y. Exploring the Limits of Language Modeling.arXiv 2016, arXiv:1602.02410.

Jun, X.U.; Tao, P.; Yao, Y. Conceptual Framework and Representation of Geographic Knowledge Map: Conceptual Framework and Representation of Geographic Knowledge Map. J. Geo-Inf. Sci. 2010, 12. [CrossRef]

Jung, C.T.; Sun, C.H. Ontology-driven problem solving framework for spatial decision support systems. Tetsu- to-Hagane. 2010, 47, 512–515.

Jung, C.T.; Sun, C.H.; Yuan, M. An ontology-enabled framework for a geospatial problem-solving environment. Comput. Environ. Urban Syst. 2013, 38, 45–57. [CrossRef]

Kang, J.; Fang, L.; Li, S.; Wang, X. Parallel Cellular Automata Markov Model for Land Use Change Prediction over MapReduce Framework. ISPRS Int. J. Geo-Inf. 2019, 8, 454. [CrossRef]

Kauppinen, T.; Espindola, G.M. Ontology-Based Modeling of Land Change Trajectories in the Brazilian Amazon. In Proceedings of the Geoinformatik, Münster, Germany, 15–17 June 2013.

Khalsa, S.J.S. Data and Metadata Brokering–Theory and Practice from the Bcube Project. Data Sci. J. 2017, 16, 1–8. [CrossRef]

Khan, M.A.; Uddin, M.F.; Gupta, N. Seven V's of Big Data understanding Big Data to extract value. In Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, Bridgeport, CT, USA, 3–5 April 2014; pp. 1–5.

Kim, K.-C.; Yun, S.-W. MR-Tree: A cache-conscious main memory spatial index structure for mobile GIS. In Proceedings of the International Workshop on Web and Wireless Geographical Information Systems, Goyang, Korea, 26–27 November 2004; pp. 167–180.

Kim, Y. Convolutional Neural Networks for Sentence Classification; Association for Computational Linguistics: Doha, Qatar, 2014.

Kliment, T.; Bordogna, G.; Frigerio, L.; Crema, A.; Boschetti, M.; Brivio, P.A.; Sterlacchini, S. Image data and metadata workflows automation in geospatial data infrastructure deployed for agricultural sector. In Proceedings of the Geoscience and Remote Sensing Symposium, Milan, Italy, 26–31 July 2015; pp. 146–149.

Koenderink, J.J.; van Doorn, A.J. Surface shape and curvature scales. Image Vis. Comput. 1992, 10, 557–564. [CrossRef]

Kuhn, W. Modeling Vs Encoding for the Semantic Web. Semant. Web 2010, 1, 11–15.

Lamprecht, A.L.; Margaria, T.; Steffen, B. Modeling and Execution of Scientific Workflows with the jABC Framework; Springer: Berlin/Heidelberg, Germany, 2014; pp. 14–29.

Lamprecht, A.L.; Steffen, B.; Margaria, T. Scientific workflows with the jabc framework. Int. J. Softw. Tools Technol. Transf. 2016, 18, 629–651. [CrossRef]

Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. Biometrics 1977, 33, 159. [CrossRef]

Lanorte, A.; De Santis, F.; Nolè, G.; Blanco, I.; Loisi, R.V.; Schettini, E.; Vox, G. Agricultural plastic waste spatial estimation by Landsat 8 satellite images. Comput. Electron. Agric. 2017, 141, 35–45. [CrossRef]

Lee, S.; Lee, D.; Lee, S. Network-oriented road map generation for unknown roads using visual images and gps-based location information. IEEE Trans. Consum. Electron. 2009, 55, 1233–1240. [CrossRef]

Lehmann, J. Dbpedia: A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. Semant. Web 2015, 6, 167–195.

Lenka, R.K.; Barik, R.K.; Gupta, N.; Ali, S.M.; Rath, A.; Dubey, H. Comparative analysis of SpatialHadoop and GeoSpark for geospatial big data analytics. In Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), Noida, India, 14–17 December 2016; pp. 484–488.

Li, D.; Li, X.; Liu, X.P.; Chen, Y.M.; Li, S.Y.; Liu, K.; Qiao, J.G.; Zheng, Y.Z.; Zhang, Y.H.; Lao, C.H. GPU-CA model for large-scale land-use change simulation. Chin. Sci. Bull. 2012, 57, 2442–2452. [CrossRef]

Li, D.R.; Cao, J.J.; Yuan, Y. Big data in smart cities. Sci. China Inf. Sci. 2015, 58, 108101. [CrossRef]

Li, J.; Liu, R.; Xiong, R. A Chinese Geographic Knowledge Base for Gir. In Proceedings of the IEEE International Conference on Computational Science and Engineering, Guangzhou, China, 21–24 July 2017.

Li, J.; Qin, Q.; Han, J.; Tang, L.-A.; Lei, K.H. Mining trajectory data and geotagged data in social media for road map inference. Trans. GIS 2014, 19, 18. [CrossRef]

Li, L.; Li, D.; Xing, X.; Yang, F.; Rong, W.; Zhu, H. Extraction of road intersections from gps traces based on the dominant orientations of roads. Int. J. Geo-Inf. 2017, 6, 403. [CrossRef]

Li, Q.; Li, D. Big data GIS. Geomat. Inf. Sci. Wuhan Univ. 2014, 39, 641–644.

Li, S.; Dragicevic, S.; Castro, F.A.; Sester, M.; Winter, S.; Coltekin, A.; Pettit, C.; Jiang, B.; Haworth, J.; Stein, A.; et al. Geospatial big data handling theory and methods: A review and research challenges. ISPRS J. Photogramm. Remote Sens. 2016, 115, 119–133. [CrossRef]

Li, S.; Dragicevic, S.; Castro, F.A.; Sester, M.; Winter, S.; Coltekin, A.; Pettit, C.; Jiang, B.; Haworth, J.; Stein, A.; et al. Geospatial big data handling theory and methods: A review and research challenges. ISPRS J. Photogramm. Remote Sens. 2016, 115, 119–133. [CrossRef]

Li, W.; Wang, S.; Bhatia, V. PolarHub: A large-scale web crawling engine for OGC service discovery in cyberinfrastructure. Comput Environ Urban Syst. 2016, 59, 195–207. [CrossRef]

Li, X. A review of the international researches on land use/land cover change. ACTA Geogr. Sin. Ed. 1996, 51, 558–565.

Li, X.; Song, J.; Huang, B. A scientific workflow management system architecture and its scheduling based on cloud service platform for manufacturing big data analytics. Int. J. Adv. Manuf. Technol. 2016, 84, 119–131. [CrossRef]

Li, X.; Yeh, A.G.-O. Neural-network-based cellular automata for simulating multiple land use changes using GIS. Int. J. Geogr. Inf. Sci. 2002, 16, 323–343. [CrossRef]

Li, X.; Zhang, X.; Yeh, A.; Liu, X. Parallel cellular automata for large-scale urban simulation using load-balancing techniques. Int. J. Geogr. Inf. Sci. 2010, 24, 803–820. [CrossRef]

Li, Z. Geospatial Big Data Handling with High Performance Computing: Current Approaches and Future Directions. In High Performance Computing for Geospatial Applications; Tang, W., Wang, S., Eds.; Springer: New York, NY, USA, 2020; ISBN 978-3-030-47997-8.

Li, Z.; Hodgson, M.E.; Li, W. A general-purpose framework for parallel processing of large-scale LiDAR data. Int. J. Digit. Earth 2018, 11, 26–47. [CrossRef]

Li, Z.; Hu, F.; Schnase, J.L.; Duffy, D.Q.; Lee, T.; Bowen, M.K.; Yang, C. A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce. Int. J. Geogr. Inf. Sci. 2017, 31, 17–35. [CrossRef]

Li, Z.; Huang, Q.; Emrich, C. Introduction to Social Sensing and Big Data Computing for Disaster Management. Int. J. Digit. Earth 2019, 12, 1198–1204. [CrossRef]

Li, Z.; Wang, C.; Emrich, C.T.; Guo, D. A novel approach to leveraging social media for rapid flood mapping: A case study of the 2015 South Carolina floods. Cartogr. Geogr. Inf. Sci. 2017, 1–14. [CrossRef]

Li, Z.; Yang, C.; Jin, B.; Yu, M.; Liu, K.; Sun, M.; Zhan, M. Enabling big geoscience data analytics with a cloud-based, MapReduce-enabled and service-oriented workflow framework. PLoS ONE 2015, 10, e0116781. [CrossRef] [PubMed]

Liang, J.; Zhong, M.; Zeng, G.; Chen, G.; Hua, S.; Li, X.; Yuan, Y.; Wu, H.; Gao, X. Risk management for optimal land use planning integrating ecosystem services values: A case study in Changsha, Middle China. Sci. Total Environ. 2017, 579, 1675–1682. [CrossRef] [PubMed]

Liang, L.; Geng, D.; Huang, T.; Di, L.; Lin, L.; Sun, Z. VCI-based Analysis of Spatio-temporal Variations of Spring Drought in China from 1981 to 2015. In Proceedings of the 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Stanbul, Turkey, 16–19 July 2019; pp. 1–6.

Lindell, M.K.; Prater, C.S.; Perry, R.W.; Nicholson, W.C. Fundamentals of Emergency Management; Emond Montgomery Publications: Toronto, ON, Canada, 2006.

Linyao, Q.; Zhiqiang, D.; Qing, Z. A task-oriented disaster information correlation method. In Proceedings of the 2015 International Workshop on Spatiotemporal Computing, Fairfax, VA, USA, 13–15 July 2015; Volume II-4/W2, pp. 169–176.

Liu, C.; Xiong, L.; Hu, X.; Shan, J. A progressive buffering method for road map update using openstreetmap data. ISPRS Int. J. Geo-Inf. 2015, 4, 1246–1264. [CrossRef]

Liu, P.; Di, L.; Du, Q.; Wang, L. Remote Sensing Big data: Theory, Methods and Applications. Remote Sens. 2018, 10, 711. [CrossRef]

Liu, X.; Biagioni, J.; Eriksson, J.; Wang, Y.; Forman, G.; Zhu, Y. Mining large-scale, sparse GPS traces for map inference: Comparison of approaches. In Proceedings of the 18th ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 669–677.

Liu, X.; Hu, G.; Chen, Y.; Li, X.; Xu, X.; Li, S.; Pei, F.; Wang, S. High-resolution multi-temporal mapping of global urban land using Landsat images based on the Google Earth Engine Platform. Remote Sens. Env. 2018, 209, 227–239. [CrossRef]

Liu, X.; Liu, X.; Wei, H.; Forman, G.; Zhu, Y. CrowdAtlas: Self-updating maps for cloud and personal use. In Proceedings of the International Conference on Mobile Systems, Applications, and Services, Taipei, Taiwan, 25–28 June 2013; pp. 469–470.

Liu, X.; Thomsen, C.; Pedersen, T.B. ETLMR: A highly scalable dimensional ETL framework based on MapReduce. In Transactions on Large-Scale Data-and Knowledge-Centered Systems VIII; Springer: Berlin/Heidelberg, Germany, 2013; pp. 1–31.

Liu, X.H.; Andersson, C. Assessing the impact of temporal dynamics on land-use change modeling. Comput. Environ. Urban Syst. 2004, 28, 107–124. [CrossRef]

Lopez, L.A.; Khalsa, S.J.S.; Duerr, R.; Tayachow, A.; Mingo, E. The BCube Crawler: Web Scale Data and Service Discovery for EarthCube. In Proceedings of the AGU Fall Meeting, San Francisco, CA, USA, 15–19 December 2014. Abstracts IN51C-06.

Lovelock, J. Gaia: The living Earth. Nature 2003, 426, 769–770. [CrossRef]

Luitjens, J.; Berzins, M.; Henderson, T. Parallel space-filling curve generation through sorting: Research Articles. Concurr. Comput. Pract. Exp. 2010, 19, 1387–1402. [CrossRef]

Luo, J. The Semantic Geospatial Problem Solving Environment: An Enabling Technology for Geographical Problem Solving under Open, Heterogeneous Environments. Ph.D. Thesis, The Pennsylvania State University, Pennsylvania, PA, USA, 2007.

Lutz, M. Ontology-based descriptions for semantic discovery and composition of geoprocessing services. Geoinformatica 2007, 11, 1–36. [CrossRef]

Mandelbrot, B.B. The Fractal Geometry of Nature; WH Freeman: New York, NY, USA, 1982; Volume 1.

Marmot from GitHub. Available online: https://github.com/kwlee0220/marmot.server.dist (accessed on 23 September 2019).

McGuire, K.J.; McDonnell, J.J.; Weiler, M.; Kendall, C.; McGlynn, B.L.; Welker, J.M.; Seibert, J. The role of topography on catchment-scale water residence time. Water Resour. Res. 2005, 41. [CrossRef]

Mei, J. From Alc to Shoq(D):A Survey of Tableau Algorithms for Description Logics. Comput. Sci. 2005, 32, 1–11. [CrossRef]

Memarian, H.; Kumar Balasundram, S.; Bin Talib, J.; Teh Boon Sung, C.; Mohd Sood, A.; Abbaspour, K. Validation of CA-Markov for Simulation of Land Use and Cover Change in the Langat Basin, Malaysia.

Mikita, T.; Balogh, P. Usage of geoprocessing services in precision forestry for wood volume calculation and wind risk assessment. Acta Univ. Agric. Silvic. Mendel. Brun. 2015, 63, 793–801. [CrossRef]

Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. arXiv 2013, arXiv:1301.3781.

Mikolov, T.; Kopecky, J.; Burget, L.; Glembek, O.; Cernocky, J. Neural network based language models for highly inflective languages. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 4725–4728.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 2013, 26, 3111–3119.

Miliaresis, G. The Landcover Impact on the Aspect/Slope Accuracy Dependence of the SRTM-1 Elevation Data for the Humboldt Range. Sensors 2008, 8, 3134–3149. [CrossRef]

Misra, S.; Saha, S.K.; Mazumdar, C. Performance Comparison of Hadoop Based Tools with Commercial ETL Tools-A Case Study. In Proceedings of the International Conference on Big Data Analytics, Mysore, India, 16–18 December 2013; pp. 176–184.

Mobaied, S.; Riera, B.; Lalanne, A.; Baguette, M.; Machon, N. The use of diachronic spatial approaches and predictive modelling to study the vegetation dynamics of a managed heathland. Biodivers. Conserv. 2011, 20, 73–88. [CrossRef]

Moore, I.D.; Grayson, R.B.; Ladson, A.R. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. Hydrol. Process. 1991, 5, 3–30. [CrossRef]

Mora, A.D.; Vieira, P.M.; Manivannan, A.; Fonseca, J.M. Automated drusen detection in retinal images using analytical modelling algorithms. Biomed. Eng. Online 2011, 10, 59. [CrossRef] [PubMed]

Morais, C.D. Where Is the Phrase "80% of Data is Geographic?". Available online: http://www.gislounge. com/80-percent-data-is-geographic (accessed on 4 April 2018).

Müller, M. Hierarchical profiling of geoprocessing services. Comput. Geosci. 2015, 82, 68–77. [CrossRef]

mundialis GmbH & Co. KG. Actinia: Geoprocessing in the Cloud. Available online: https://actinia.mundialis. de/ (accessed on 10 March 2020).

Narayanan, V.; Arora, I.; Bhatia, A. Fast and accurate sentiment classification using an enhanced Naive Bayes model. In Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning; Springer: Berlin/Heidelberg, Germany, 2013; pp. 194–201.

Neria, Y.; Shultz, J.M. Mental health effects of Hurricane Sandy: Characteristics, potential aftermath, and response. J. Am. Med. Assoc. 2012, 308, 2571–2572. [CrossRef] [PubMed]

Niehöfer, B.; Burda, R.; Wietfeld, C.; Bauer, F.; Lueert, O. GPS Community Map Generation for Enhanced Routing Methods Based on Trace-Collection by Mobile Phones. In Proceedings of the International Conference on Advances in Satellite and Space Communications, Colmar, France, 20–25 July 2009; pp. 156–161.

Nogueras, J.; Zarazaga, F.J.; Muro, R.P. Interoperability between metadata standards. In Geographic Information Metadata for Spatial Data Infrastructures; Springer: Berlin/Heidelberg, Germany, 2005; pp. 89–127.

Nogueras-Iso, J.; Zarazaga-Soria, F.J.J.; Be´jarbe´jar, R.; Lvarez, P.J.A.; Muro-Medrano, P.R.R.; Béjar, R.; Álvarez, P.J.; Muro-Medrano, P.R.R. OGC Catalog Services: A key element for the development of Spatial Data Infrastructures. Comput. Geosci. 2005, 31, 199–209. [CrossRef]

Norris, F.H.; Friedman, M.J.; Watson, P.J.; Byrne, C.M.; Diaz, E.; Kaniasty, K. 60,000 Disaster Victims Speak: Part I. An Empirical Review of the Empirical Literature, 1981–2001. Psychiatry-Interpers. Biol. Process. 2002, 65, 207–239. [CrossRef]

OGC. OGC Abstract Specifications: Topic 5—Features; Open Geospatial Consortium: Wayland, MA, USA, 2009; pp. 8–126.

Oh, O.; Kwon, K.H.; Rao, H.R. An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter during the Haiti Earthquake 2010. In Proceedings of the International Conference on Information Systems, Icis 2010, Saint Louis, MO, USA, 12–15 December 2010; p. 231.

Olaya, V.; Conrad, O. Geomorphometry in SAGA. In Developments in Soil Science; Elsevier: Amsterdam, The Netherlands, 2009; Volume 33, pp. 293–308.

Oliveira, S.; Pereira, J.M.C.; San-Miguel-Ayanz, J.; Lourenço, L. Exploring the spatial patterns of fire density in Southern Europe using Geographically Weighted Regression. Appl. Geogr. 2014, 51, 143–157. [CrossRef]

Pallickara, S.L.; Pallickara, S.; Zupanski, M.; Sullivan, S. Efficient Metadata Generation to Enable Interactive Data Discovery over Large-scale Scientific Data Collections. In Proceedings of the 2010 IEEE Second International Conference on Cloud Computing Technology and Science, Indianapolis, IN, USA, 30 November–3 December 2010.

Pang, T.B.; Pang, B.; Lee, L. Thumbs up? Sentiment Classification using Machine Learning. Empir. Methods Nat. Lang. Process. 2002, 10, 79–86.

Pappas, N.; Popescu-Belis, A. Multilingual Hierarchical Attention Networks for Document Classification. arXiv 2017, arXiv:1707.00896.

Park, H.; Yoon, A.; Kwon, H.C. Task model and task ontology for intelligent tourist information service.

Park, S.; Bang, Y.; Yu, K. Techniques for updating pedestrian network data including facilities and obstructions information for transportation of vulnerable people. Sensors 2015, 15, 24466–24486. [CrossRef]

Peitgen, H.-O.; Jürgens, H.; Saupe, D. Chaos and Fractals: New Frontiers of Science; Springer Science & Business Media: New York, NY, USA, 1992.

Peng, Y.; Gong, J.; Di, L.; Jie, Y.; Sun, L.; Sun, Z.; Qian, W. GeoPW: Laying blocks for the geospatial processing web. Trans. GIS 2010, 14, 755–772.

Perez, A.G.; Benjamins, V.R. Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods. In Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99), Stockholm, Sweden, 31 July–6 August 1999.

Pike, R.J. Geomorphometry-diversity in quantitative surface analysis. Prog. Phys. Geogr. Earth Env. 2000, 24, 1–20. [CrossRef]

Pittet, P.; Cruz, C.; Nicolle, C. Modeling Changes for Shoin(D) Ontologies: An Exhaustive Structural Model. In Proceedings of the IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, 16–18 September 2013.

Puri, S.; Prasad, S.K. Efficient parallel and distributed algorithms for GIS polygonal overlay processing. In Proceedings of the 2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum, Cambridge, MA, USA, 20–24 May 2013; pp. 2238–2241.

Qi, K.; Gui, Z.; Li, Z.; Guo, W.; Wu, H.; Gong, J. An extension mechanism to verify, constrain and enhance geoprocessing workflows invocation. Trans. GIS 2016, 20, 240–258. [CrossRef]

Qiu, B.; Chen, C. Land use change simulation model based on MCDM and CA and its application. Acta Geogr. Sin. Ed. 2008, 63, 165–174.

Qiu, J.; Wang, R. Automatic extraction of road networks from gps traces. Photogramm. Eng. Remote Sens. 2016, 82, 593–604. [CrossRef]

Qu, Y.; Huang, C.; Zhang, P.; Zhang, J. Microblogging after a major disaster in China: A case study of the 2010 Yushu earthquake. In Proceedings of the ACM 2011 Conference on

Computer Supported Cooperative Work, Hangzhou, China, 19–23 March 2010; pp. 25–34.

Rahman, M.T.U.; Tabassum, F.; Rasheduzzaman, M.; Saba, H.; Sarkar, L.; Ferdous, J.; Uddin, S.Z.; Islam, A.Z.M.Z. Temporal dynamics of land use/land cover change and its prediction using CA-ANN model for southwestern coastal Bangladesh. Environ. Monit. Assess. 2017, 189, 565. [CrossRef] [PubMed]

Rao, J.; Wu, B.; Dong, Y.-X. Parallel Link Prediction in Complex Network Using MapReduce. J. Softw. 2014, 23, 3175–3186. [CrossRef]

Rathore, M.M.U.; Paul, A.; Ahmad, A.; Chen, B.; Huang, B.; Ji, W. Real-Time Big Data Analytical Architecture for Remote Sensing Application. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2015, 8, 4610–4621. [CrossRef]

Reid, W.V.; Bréchignac, C.; Tseh Lee, Y. Earth system research priorities. Science 2009, 325, 245. [CrossRef] [PubMed]

Report on Text Classification Using CNN, RNN & HAN. Available online: https://medium.com/jatana/ report-on-text-classification-using-cnn-rnn-han-f0e887214d5f (accessed on 12 January 2019).

Rimal, B.; Zhang, L.; Keshtkar, H.; Wang, N.; Lin, Y. Monitoring and Modeling of Spatiotemporal Urban Expansion and Land-Use/Land-Cover Change Using Integrated Markov Chain Cellular Automata Model. ISPRS Int. J. Geo. Inf. 2017, 6, 288. [CrossRef]

Rossignac, J. Shape complexity. Vis. Comput. 2005, 21, 985–996. [CrossRef]

Saaty, T.L.; Vargas, L.G. Models, Methods, Concepts & Applications of the Analytic Hierarchy Process; Springer Science & Business Media: New York, NY, USA, 2001; ISBN 978-1-4613-5667-7.

Sabtu, A.; Azmi, N.F.M.; Sjarif, N.N.A.; Ismail, S.A.; Yusop, O.M.; Sarkan, H.; Chuprat, S. The challenges of extract, transform and loading (ETL) system implementation for near real-time environment. In Proceedings of the 2017 International Conference on Research and Innovation in Information Systems (ICRIIS) 2017, Langkawi, Malaysia, 16–17 July 2017; pp. 1–5.

Safanelli, J.L.; Poppiel, R.R.; Ruiz, L.F.C.; Bonfatti, B.R.; Mello, F.A.d.O.; Rizzo, R.; Demattê, J.A.M. Terrain

Samadzadegan, F.; Saber, M.; Zahmatkesh, H.; Joze Ghazi Khanlou, H. An architecture for automated fire detection early warning system based on geoprocessing service composition. In Proceedings of the SMPR 2013, Tehran, Iran, 5–8 October 2013; pp. 351–355.

Sankey, T.T.; McVay, J.; Swetnam, T.L.; McClaran, M.P.; Heilman, P.; Nichols, M. UAV hyperspectral and lidar data and their fusion for arid and semi-arid land vegetation monitoring. Remote Sens. Ecol. Conserv. 2018, 4, 20–33. [CrossRef]

Sattler, U.; Horrocks, I. A description logic with transitive and inverse roles and role hierarchies. J. Log. Comput. 1999, 9, 385–410. [CrossRef]

Schellnhuber, H.J. 'Earth system' analysis and the second Copernican revolution. Nature 1999, 402, C19–C23. [CrossRef]

Schnase, J.L.; Duffy, D.Q.; Tamkin, G.S.; Nadeau, D.; Thompson, J.H.; Grieg, C.M.; McInerney, M.A.; Webster, W.P. MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service. Comput. Environ. Urban Syst. 2017, 61, 198–211. [CrossRef]

Schroedl, S.; Wagstaff, K.; Rogers, S.; Langley, P.; Wilson, C. Mining gps traces for map refinement. Data Min. Knowl. Discov. 2004, 9, 59–87. [CrossRef]

Serneels, S.; Lambin, E.F. Proximate causes of land-use change in Narok district, Kenya: A spatial statistical model. Agric. Ecosyst. Environ. 2001, 85, 65–81. [CrossRef]

Sghaier, M.O.; Lepage, R. Road Extraction From Very High Resolution Remote Sensing Optical Images Based on Texture Analysis and Beamlet Transform. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2016, 9, 1946–1958. [CrossRef]

Shan, Z.; Wu, H.; Sun, W.; Zheng, B. COBWEB: A robust map update system using GPS trajectories. In Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing, Osaka, Japan, 7–11 September 2015; pp. 927–937.

She, J.; Feng, X.; Liu, B.; Xiao, P.; Wang, P. Conceptual Data Modeling on the Evolution of the Spatiotemporal Object; Chen, J., Pu, Y., Eds.; International Society for Optics and Photonics: The Hague, The Netherlands, 2007; Volume 6753, p. 67530H.

Sherretz, L.A.; Fulker, D.W. Unidata: Enabling Universities to Acquire and Analyze Scientific Data. Bull. Am. Meteorol. Soc. 1988, 69, 373–376. [CrossRef]

Shi, W.; Shen, S.; Liu, Y. Automatic generation of road network map from massive GPS, vehicle trajectories. In Proceedings of the International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO, USA, 4–7 October 2009; pp. 1–6.

Shook, E.; Bowlick, F.J.; Kemp, K.K.; Ahlqvist, O.; Carbajeles-Dale, P.; Di Biase, D.; Rush, J. Cyber literacy for GIScience: Toward formalizing geospatial computing education. Prof. Geogr. 2019, 71, 221–238. [CrossRef]

Shukla, J.; Palmer, T.N.; Hagedorn, R.; Hoskins, B.; Kinter, J.; Marotzke, J.; Miller, M.; Slingo, J.; Shukla, J.; Palmer, T.N.; et al. Toward a New Generation of World Climate Research and Computing Facilities. Bull. Am. Meteorol. Soc. 2010, 91, 1407–1412. [CrossRef]

Silverman, B.W. Density Estimation for Statistics and Data Analysis; Chapman and Hall: San Francisco, CA, USA, 1986; pp. 296–297.

Simmhan, Y.L.; Plale, B.; Gannon, D. A survey of data provenance in e-science. ACM SIGMOD Rec. 2005, 34, 31–36. [CrossRef]

Singh, G.; Bharathi, S.; Chervenak, A.; Deelman, E.; Kesselman, C.; Manohar, M.; Patil, S.; Pearlman, L. A Metadata Catalog Service for Data Intensive Applications. In Proceedings of the 2003 ACM/IEEE Conference on Supercomputing, Phoenix, AZ, USA, 15–21 November 2003.

Singhal, A. Official Google Blog: Introducing the Knowledge Graph: Things, Not Strings; Northwestern University: Evanston, IL, USA, 2012.

Siricharoen, W.V.; Pakdeetrakulwong, U. A Survey on Ontology-Driven Geographic Information Systems. In Proceedings of the Fourth International Conference on Digital Information and Communication Technology and It's Applications, Bangkok, Thailand, 6–8 May 2014.

Social Media in Disasters and Emergencies; The Drum: Washington, DC, USA, 3 March 2013.

Song, J.; Di, L. Near-Real-Time OGC Catalogue Service for Geoscience Big Data. ISPRS Int. J. Geo-Inf. 2017, 6, 337. [CrossRef]

Spéry, L.; Claramunt, C.; Libourel, T. A Spatio-Temporal Model for the Manipulation of Lineage Metadata. Geoinformatica 2001, 5, 51–70. [CrossRef]

Storey, V.C.; Song, I.Y. Big data technologies and management: What conceptual modeling can do. Data Knowl. Eng. 2017, 108, 50–67. [CrossRef]

Šuba, R.; Meijers, M.; Oosterom, P.V. Continuous road network generalization throughout all scales. ISPRS Int. J. Geo-Inf. 2016, 5, 145. [CrossRef]

Subedi, P.; Subedi, K.; Thapa, B. Application of a Hybrid Cellular Automaton—Markov (CA-Markov) Model in Land-Use Change Prediction: A Case Study of Saddle Creek Drainage Basin, Florida. Appl. Ecol. Environ. Sci. 2013, 1, 126–132.

Suchanek, F.M.; Kasneci, G.; Weikum, G. Yago: A Core of Semantic Knowledge. In Proceedings of the 16th International Conference on World Wide Web (WWW), Banff, AB, Canada, 8–12 May 2007; Volume 272, pp. 697–706.

Sun, Z.; Di, L. CyberConnector COVALI: Enabling inter-comparison and validation of Earth science models. In Proceedings of the AGU Fall Meeting, Washington, DC, USA, 10–14 December 2018. Abstract #IN23B-0780.

Sun, Z.; Di, L.; Gaigalas, J. SUIS: Simplify the use of geospatial web services in environmental modelling. Environ. Model. Softw. 2019, 119, 228–241. [CrossRef]

Sun, Z.; Di, L.; Hao, H.; Wu, X.; Tong, D.Q.; Zhang, C.; Virgei, C.; Fang, H.; Yu, E.; Tan, X.; et al. CyberConnector: A service-oriented system for automatically tailoring multisource Earth observation data to feed Earth science models. Earth Sci. Inform. 2018, 11, 1–17. [CrossRef]

Sun, Z.; Di, L.; Zhang, C.; Fang, H.; Yu, E.; Lin, L.; Tang, J.; Tan, X.; Liu, Z.; Jiang, L.; et al. Building robust geospatial web services for agricultural information extraction and sharing. In Proceedings of the 2017 6th International Conference on Agro-Geoinformatics, Fairfax, VA, USA, 7–10 August 2017; pp. 1–4.

Sun, Z.; Peng, C.; Deng, M.; Chen, A.; Yue, P.; Fang, H.; Di, L. Automation of Customized and Near-Real-Time Vegetation Condition Index Generation Through Cyberinfrastructure-Based Geoprocessing Workflows. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2014, 7, 4512–4522. [CrossRef]

Sun, Z.; Yue, P.; Di, L. Geopwtmanager: A task-oriented web geoprocessing system. Comput. Geosci. 2012, 47, 34–45. [CrossRef]

Sun, Z.; Yue, P.; Di, L. GeoPWTManager: A task-oriented web geoprocessing system. Comput. Geosci. 2012, 47, 34–45. [CrossRef]

Sun, Z.; Yue, P.; Hu, L.; Gong, J.; Zhang, L.; Lu, X. GeoPWProv: Interleaving Map and Faceted Metadata for Provenance Visualization and Navigation. IEEE Trans. Geosci. Remote Sens. 2013, 51, 5131–5136.

Sun, Z.; Yue, P.; Lu, X.; Zhai, X.; Hu, L. A task ontology driven approach for live geoprocessing in a service-oriented environment. Trans. GIS 2012, 16, 867–884. [CrossRef]

Sun, Z.; Yue, P.; Lu, X.; Zhai, X.; Hu, L. A Task Ontology Driven Approach for Live Geoprocessing in a Service-Oriented Environment. Trans. GIS 2012, 16, 867–884. [CrossRef]

Tamiminia, H.; Salehi, B.; Mahdianpari, M.; Quackenbush, L.; Adeli, S.; Brisco, B. Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. ISPRS J. Photogramm. Remote Sens. 2020, 164, 152–170. [CrossRef]

Tan, X.; Di, L.; Deng, M.; Chen, A.; Sun, Z.; Huang, C.; Shao, Y.; Ye, X. Agent-and Cloud-Supported Geospatial Service Aggregation for Flood Response. ISPRS Ann Photogramm. Remote Sens. Spat. Inf. Sci. 2015, 2, 13–18. [CrossRef]

Tan, X.; Di, L.; Deng, M.; Fu, J.; Shao, G.; Gao, M.; Sun, Z.; Ye, X.; Sha, Z.; Jin, B. Building an Elastic Parallel OGC Web Processing Service on a Cloud-Based Cluster: A Case Study of Remote Sensing Data Processing Service. Sustainability 2015, 7, 14245–14258. [CrossRef]

Tan, X.; Di, L.; Deng, M.; Huang, F.; Ye, X.; Sha, Z.; Sun, Z.; Gong, W.; Shao, Y.; Huang, C. Agent-as-a-service-based geospatial service aggregation in the cloud: A case study of flood response. Environ. Model. Softw. 2016, 84, 210–225. [CrossRef]

Tang, D.; Qin, B.; Liu, T. Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1422–1432.

Tausczik, Y.R.; Pennebaker, J.W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. J. Lang. Soc. Psychol. 2009, 29, 24–54. [CrossRef]

Taylor, I.; Shields, M.; Wang, I.; Harrison, A. The triana workflow environment: Architecture and applications. In Workflows e-Science; Springer: London, UK, 2007; pp. 320–339.

The International Organization for Standardization (ISO). ISO 19107: Geographic Information—Spatial Schema; The International Organization for Standardization: Geneva, Switzerland, 2003.

Theodoridis, Y.; Sellis, T.; Papadopoulos, A.N.; Manolopoulos, Y. Specifications for Efficient Indexing in Spatiotemporal Databases. In Proceedings of the Tenth International Conference on Scientific and Statistical Database Management, Capri, Italy, 3 July 1998.

Thomsen, C.; Bach Pedersen, T. pygrametl: A powerful programming framework for extract-transform-load programmers. In DOLAP '09, Proceedings of the ACM Twelfth International Workshop on Data Warehousing and OLAP, Hong Kong, China, 6 November 2009; ACM: New York, NY, USA, 2009; pp. 49–56.

Tilove, R.B. Line/polygon classification: A study of the complexity of geometric computation. IEEE Comput. Graph. Appl. 1981, 1, 75–88. [CrossRef]

Tran, V.X.; Tsuji, H. Owl-t: An ontology-based task template language for modeling business processes. In Proceedings of the Acis International Conference on Software Engineering Research, Management & Applications, Busan, Korea, 20–22 August 2007; pp. 101–108.

Trujillo, J.; Lujan-Mora, S. A UML based approach for modeling ETL processes in data warehouses. In Conceptual Modeling—ER 2003, Proceedings of the International Conference on Conceptual Modeling, Chicago, IL, USA, 13–16 October 2003; Springer: Berlin/Heidelberg, Germany, 2003; pp. 307–320.

Turney, P.D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, 6 July 2012; pp. 417–424.

Unidata THREDDs Client Catalog Spec 1.0.7. Available online: https://www.unidata.ucar.edu/software/tds/ current/catalog/InvCatalogSpec.html (accessed on 26 August 2019).

Unidata THREDDS Support [THREDDS #BIA-775104]: Unidata THREDDs Metadata Structure and Volume. Juozasgaigalas@gmail.com. Gmail. Available online:

https://mail.google.com/mail/u/0/#search/Unidata+THREDDs+metadata+structure+and+volume/FMfcgxvwzcCgSZmpPZsQFqdjLlCkPNfm (accessed on 26 August 2019).

USGS EROS. GTOPO30-Global 1-km Digital Raster Data Derived from a Variety of Sources. Available online: https://doi.org/10.5066/F7DF6PQS (accessed on 10 March 2020).

USGS EROS. USGS EROS Archive-Digital Elevation-Shuttle Radar Topography Mission (SRTM) Void Filled. Available online: https://doi.org/10.5066/F7F76B1X (accessed on 4 April 2020).

Vahedi, B.; Kuhn, W.; Ballatore, A. Question-based spatial computing—A case study. In Geospatial Data in a Changing World; Springer International Publishing: Cham, Switzerland, 2016; pp. 37–50.

van Kreveld, M.; Nievergelt, J.; Roos, T.; Widmayer, P. Algorithmic Foundations of Geographic Information Systems; Springer: New York, NY, USA, 1997; Volume 1340.

van Vliet, J.; Bregt, A.K.; Hagen-Zanker, A. Revisiting Kappa to account for change in the accuracy assessment of land-use change models. Ecol. Model. 2011, 222, 1367–1375. [CrossRef]

Vatti, B.R. A generic solution to polygon clipping. Commun. ACM 1992, 35, 56–63. [CrossRef]

Veldkamp, A.; Lambin, E. Predicting land-use change. Agric. Ecosyst. Environ. 2001, 85, 1–6. [CrossRef]

Verburg, P.H.; Soepboer, W.; Veldkamp, A.; Limpiada, R.; Espaldon, V.; Mastura, S.S.A. Modeling the spatial dynamics of regional land use: The CLUE-S model. Environ. Manag. 2002, 30, 391–405. [CrossRef] [PubMed]

Viera, A.J.; Garrett, J.M. Understanding interobserver agreement: The kappa statistic. Fam. Med. 2005, 37, 360–363.

Wang, H.; Chong, S. A high efficient polygon clipping algorithm for dealing with intersection degradation. J. Southeast Univ. 2016, 4, 702–707.

Wang, J. An Efficient Algorithm for Complex Polygon Clipping. Geomat. Inf. Sci. Wuhan Univ. 2010, 35, 369–372.

Wang, S.; Zhang, X.; Ye, P.; Du, M.; Lu, Y.; Xue, H. Geographic Knowledge Graph (GeoKG): A Formalized Geographic Knowledge Representation. ISPRS Int. J. Geo-Inf. 2019, 8, 184. [CrossRef]

Wang, S.; Zhong, E.; Lu, H.; Guo, H.; Long, L. An effective algorithm for lines and polygons overlay analysis using uniform spatial grid indexing. In Proceedings of the 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), Fuzhou, China, 8–10 July 2015; pp. 175–179.

Wang, Y.; Feng, S.; Wang, D.; Yu, G.; Zhang, Y. Multi-label Chinese Microblog Emotion Classification via Convolutional Neural Network. In Proceedings of the Web Technologies and Applications: 18th Asia-Pacific Web Conference, APWeb 2016, Suzhou, China, 23–25 September 2016.

Wang, Y.; Liu, Z.; Liao, H.; Li, C. Improving the performance of GIS polygon overlay computation with MapReduce for spatial big data processing. Clust. Comput. 2015, 18, 507–516. [CrossRef]

Wang, Y.; Wei, H.; Forman, G. Mining large-scale gps streams for connectivity refinement of road maps. Comput. J. 2018, 58, 2109–2119.

Wei, Y.; Di, L.; Zhao, B.; Liao, G.; Chen, A. Transformation of HDF-EOS metadata from the ECS model to ISO 19115-based XML. Comput. Geosci. 2007, 33, 238–247. [CrossRef]

Weiler, K.; Atherton, P. Hidden surface removal using polygon area sorting. ACM SIGGRAPH Comput. Graph. 1977, 11, 214–222. [CrossRef]

West, L.A.; Hess, T.J. Metadata as a knowledge management tool: Supporting intelligent agent and end user access to spatial data. Decis. Support Syst. 2002, 32, 247–264. [CrossRef]

White, T. Hadoop: The Definitive Guide, 3rd ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2012; ISBN 1449338771.

White, R.; Engelen, G. Cellular Automata and Fractal Urban Form: A Cellular Modelling Approach to the Evolution of Urban Land-Use Patterns. Environ. Plan. A Econ. Spec. 1993, 25, 1175–1199. [CrossRef]

White, R.; Engelen, G.; Uljee, I. The use of constrained cellular automata for high-resolution modelling of urban land-use dynamics. Environ. Plan. B Plan. Des. 1997, 24, 323–343. [CrossRef]

Wiegand, N.; García, C. A task-based ontology approach to automate geospatial data retrieval. Trans. GIS 2007, 11, 355–376. [CrossRef]

Wijesekara, G.N.; Farjad, B.; Gupta, A.; Qiao, Y.; Delaney, P.; Marceau, D.J. A comprehensive land-use/hydrological modeling system for scenario simulations in the Elbow River watershed, Alberta, Canada. Environ. Manag. 2014, 53, 357–381. [CrossRef] [PubMed]

Wiley, K.; Connolly, A. Astronomical image processing with hadoop. Astron. Data Anal. Softw. Syst. 2010, 442, 93–96.

William, M. (Ed.) The American Heritage Dictionary of the English Language; New College Edition; Houghton Mifflin Company: Boston, MA, USA, 1980.

Winden, K.V.; Biljecki, F.; Spek, S.V.D. Automatic update of road attributes by mining gps tracks. Trans. GIS 2016, 20, 664–683. [CrossRef]

Wolfram, S. Cellular automata as models of complexity. Nature 1984, 311, 419–424. [CrossRef]

Wolstencroft, K.; Haines, R.; Fellows, D.; Williams, A.; Withers, D.; Owen, S.; Soilandreyes, S.; Dunlop, I.; Nenadic, A.; Fisher, P. The taverna workflow suite: Designing and executing workflows of web services on the desktop, web or in the cloud. Nucleic Acids Res. 2013, 41, 557–561. [CrossRef] [PubMed]

Word2Vec. Available online: https://code.google.com/archive/p/word2vec/ (accessed on 12 January 2019).

Wright, D.J.; Wang, S. The emergence of spatial cyberinfrastructure. Proc. Natl. Acad. Sci. USA 2011, 108, 5488–5491. [CrossRef] [PubMed]

Wu, H.; Xu, Z.; Wu, G. A Novel Method of Missing Road Generation in City Blocks Based on Big Mobile Navigation Trajectory Data. ISPRS Int. J. Geo-Inf. 2019, 8, 142. [CrossRef]

Wu, T.; Xiang, L.; Gong, J. Updating road networks by local renewal from gps trajectories. ISPRS Int. J. Geo-Inf. 2016, 5, 163. [CrossRef]

Wu, W.; Li, H.; Wang, H.; Zhu, K.Q. Probase: A Probabilistic Taxonomy for Text Understanding. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD'12), Scottsdale, AZ, USA, 20–24 May 2012.

Wu, X.; Liu, X.; Zhou, S. Principle and Method of MapGIS IGServer; Publishing House of Electronics Industry: Beijing, China, 2012.

Xia, J.; Yang, C.; Li, Q. Building a spatiotemporal index for Earth Observation Big Data. Int. J. Appl. Earth Obs. Geoinf. 2018, 73, 245–252. [CrossRef]

Xiao, Z.; Li, X.; Wang, L.; Yang, Q.; Du, J.; Sangaiah, A.K. Using convolution control block for Chinese sentiment analysis. J. Parallel Distrib. Comput. 2017, 116, 18–26. [CrossRef]

Xiao, Z.; Qiu, Q.; Fang, J.; Cui, S. A vector map overlay algorithm based on distributed queue. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 6098–6101.

Xie, X.; Bingyungwong, K.; Aghajan, H.; Veelaert, P.; Philips, W. Inferring directed road networks from gps traces by track alignment. ISPRS Int. J. Geo-Inf. 2015, 4, 2446–2471. [CrossRef]

Xiong, F.; Deng, Y.; Tang, X. The Architecture of Word2vec and Its Applications. J. Nanjing Norm. Univ. 2015, 1, 43–48.

Xu, R.; Wong, K.F.; Xia, Y. Coarse-Fine Opinion Mining—WIA in NTCIR-7 MOAT Task. In Proceedings of the NTCIR 2008, Tokyo, Japan, 16–19 December 2008.

Yang, C.; Goodchild, M.; Gahegan, M. Geospatial Cyberinfrastructure: Past, present and future. Comput. Environ. Urban Syst. 2010, 34, 264–277. [CrossRef]

Yang, C.; Goodchild, M.; Huang, Q.; Nebert, D.; Raskin, R.; Xu, Y.; Bambacus, M.; Fay, D. Spatial cloud computing: How can the geospatial sciences use and help shape cloud computing? Int. J. Digit. Earth 2011, 4, 305–329. [CrossRef]

Yang, C.; Wu, H.; Huang, Q.; Li, Z.; Li, J. Using spatial principles to optimize distributed computing for enabling the physical science discoveries. Proc. Natl. Acad. Sci. USA 2011, 108, 5498–5503. [CrossRef] [PubMed]

Yang, Q.; Li, X.; Shi, X. Cellular automata for simulating land use changes based on support vector machines. Comput. Geosci. 2008, 34, 592–602. [CrossRef]

Yang, T.; Xie, J.; Li, G.; Mou, N.; Li, Z.; Tian, C.; Zhao, J. Social Media Big Data Mining and Spatio-Temporal Analysis on Public Emotions for Disaster Mitigation. ISPRS Int. J. Geo-Inf. 2019, 8, 29. [CrossRef]

Yang, X.; Zheng, X.-Q.; Lv, L.-N. A spatiotemporal model of land use change based on ant colony optimization, Markov chain and cellular automata. Ecol. Model. 2012, 233, 11–19. [CrossRef]

Yang, Z.L.; Cao, J.; Hu, K.; Gui, Z.P.; Wu, H.Y.; You, L. Developing a cloud-based online geospatial information sharing and geoprocessing platform to facilitate collaborative education and research. In Proceedings of the ISPRS—International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016 XXIII ISPRS Congress, Prague, Czech Republic, 12–19 July 2016; Volume XLI-B6, pp. 3–7.

Yao, X.; Mokbel, M.; Ye, S.; Li, G.; Alarabi, L.; Eldawy, A.; Zhao, Z.; Zhao, L.; Zhu, D. LandQv2: A MapReduce- based system for processing arable land quality big data. ISPRS Int. J. Geo-Inf. 2018, 7, 271. [CrossRef]

Yi, J.; Nasukawa, T.; Bunescu, R.; Niblack, W. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. In Proceedings of the IEEE International Conference on Data Mining, Melbourne, FL, USA, 22 November 2003; pp. 427–434.

Yin, W.; Kann, K.; Yu, M.; Schtze, H. Comparative Study of CNN and RNN for Natural Language Processing. arXiv 2017, arXiv:1702.01923.

Ying, F.; Mooney, P.; Corcoran, P.; Winstanley, A.C. A model for progressive transmission of spatial data based on shape complexity. Sigspat. Spec. 2010, 2, 25–30. [CrossRef]

Yu, J.; Wu, J.; Sarwat, M. A demonstration of GeoSpark: A cluster computing framework for processing big spatial data. In Proceedings of the 2016 IEEE 32nd International Conference on Data Engineering (ICDE), Helsinki, Finland, 16–20 May 2016; pp. 1410–1413.

Yu, J.; Wu, J.; Sarwat, M. Geospark: A cluster computing framework for processing large-scale spatial data. In Proceedings of the 23rd SIGSPATIAL International Conference on

Advances in Geographic Information Systems, Bellevue, WA, USA, 3–6 November 2015; p. 70.

Yu, J.; Wu, J.; Sarwat, M. Geospark: A cluster computing framework for processing large-scale spatial data. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Bellevue, WA, USA, 3–6 November 2015; p. 70.

Yu, J.; Wu, J.; Sarwat, M. GeoSpark: A cluster computing framework for processing large-scale spatial data. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 3–6 November 2015; pp. 1–4.

Yu, J.; Zhang, Z.; Sarwat, M. Spatial data management in apache spark: The geospark perspective and beyond. Geoinformatica 2019, 23, 37–78. [CrossRef]

Yuan, X.; Liu, G. A task ontology model for domain independent dialogue management. In Proceedings of the IEEE International Conference on Virtual Environments Human-Computer Interfaces and Measurement Systems, Tianjin, China, 2–4 July 2012; pp. 148–153.

Yue, P.; Baumann, P.; Bugbee, K.; Jiang, L. Towards intelligent giservices. Earth Sci. Inform. 2015, 8, 463–481. [CrossRef]

Yue, P.; Di, L.; Yang, W.; Yu, G.; Zhao, P. Semantics-based automatic composition of geospatial web service chains. Comput. Geosci. 2007, 33, 649–665. [CrossRef]

Yue, P.; Gong, J.; Di, L. Augmenting geospatial data provenance through metadata tracking in geospatial service chaining. Comput. Geosci. 2010, 36, 270–281. [CrossRef]

Yue, P.; Gong, J.; Di, L.; Yuan, J.; Sun, L.; Sun, Z.; Wang, Q. GeoPW: Laying Blocks for the Geospatial Processing Web. Trans. GIS 2010, 14, 755–772. [CrossRef]

Yue, P.; Sun, Z.; Gong, J.; Di, L.; Lu, X. A provenance framework for Web geoprocessing workflows. In Proceedings of the 2011 IEEE International Geoscience and Remote Sensing Symposium, Vancouver, BC, Canada, 24–29 July 2011; pp. 3811–3814.

Yue, P.; Zhang, M.; Tan, Z. A geoprocessing workflow system for environmental monitoring and integrated modelling. Environ. Model. Softw. 2015, 69, 128–140. [CrossRef]

Zaharia, M.; Chowdhury, M.; Franklin, M.J.; Shenker, S.; Stoica, I. Spark: Cluster Computing with Working Sets; HotCloud: Boston, MA, USA, 22 June 2010.

Zaharia, M.; Xin, R.S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Ghodsi, A. Apache Spark: A unified engine for big data processing. Commun. ACM 2016, 59, 56–65. [CrossRef]

Zhang, C.; Di, L.; Sun, Z.; Lin, L.; Yu, E.G.; Gaigalas, J. Exploring cloud-based Web Processing Service: A case study on the implementation of CMAQ as a Service. Environ. Model. Softw. 2019, 113, 29–41. [CrossRef]

Zhang, C.; Li, W. Markov chain modeling of multinomial land-cover classes. GISci. Remote Sens. 2005, 42, 1–18. [CrossRef]

Zhang, C.; Zhao, T.; Li, W. Automatic search of geospatial features for disaster and emergency management. Int. J. Appl. Earth Obs. Geoinf. 2010, 12, 409–418. [CrossRef]

Zhang, D.; Wang, D. Relation Classification: CNN or RNN? 2016. Available online: https://link.springer. com/chapter/10.1007/978-3-319-50496-4_60 (accessed on 12 January 2019).

Zhang, J.; Wang, Y.; Zhao, W. An Improved Hybrid Method for Enhanced Road Feature Selection in Map Generalization. ISPRS Int. J. Geo-Inf. 2017, 6, 196. [CrossRef]

Zhang, L.; Thiemann, F.; Sester, M. Integration of GPS traces with road map. In Proceedings of the Second International Workshop on Computational Transportation Science, Seattle, WA, USA, 3 November 2009; pp. 17–22.

Zhang, M.; Bu, X.; Yue, P. Geojmodelbuilder: An open source geoprocessing workflow tool. Open Geospat. Data Softw. Stand. 2017, 2, 8. [CrossRef]

Zhang, S.Q.; Zhang, C.; Yang, D.H.; Zhang, J.Y.; Pan, X.; Jiang, C.L. Overlay of Polygon Objects and Its Parallel Computational Strategies Using Simple Data Model. Geogr. Geo-Inf. Sci. 2013, 29, 43–46.

Zhang, T.; Wang, J.; Cui, C.; Li, Y.; He, W.; Lu, Y.; Qiao, Q. Integrating Geovisual Analytics with Machine Learning for Human Mobility Pattern Discovery. ISPRS Int. J. Geo-Inf. 2019, 8, 434. [CrossRef]

Zhang, Y.; Gao, Y.; Xue, L.L.; Shen, S.; Chen, K. A common sense geographic knowledge base for GIR. Sci. Technol. Sci. 2008, 51, 26–37. [CrossRef]

Zhang, Y.; Liu, J.; Qian, X.; Qiu, A.; Zhang, F. An Automatic Road Network Construction Method Using Massive GPS Trajectory Data. ISPRS Int. J. Geo-Inf. 2017, 6, 400. [CrossRef]

Zhao, K.; Jin, B.; Fan, H.; Song, W.; Zhou, S.; Jiang, Y. High-Performance Overlay Analysis of Massive Geographic Polygons That Considers Shape Complexity in a Cloud Environment. ISPRS Int. J. Geo-Inf. 2019, 8, 290. [CrossRef]

Zhao, P. Geospatial Web Services: Advances in Information Interoperability: Advances in Information Interoperability; IGI Global: Hershey, PA, USA, 2010.

Zhao, P.; Di, L.; Yu, G.; Yue, P.; Wei, Y.; Yang, W. Semantic web-based geospatial knowledge transformation. Comput. Geosci. 2009, 35, 798–808. [CrossRef]

Zhao, P.; Yu, G.; Di, L. Geospatial Web Services. In Emerging Spatial Information Systems and Applications, 1st ed.; IGI Global: Hershey, PA, USA, 2006; pp. 1–35.

Zhao, Y.; Liu, J.; Chen, R.; Li, J.; Xie, C.; Niu, W. A new method of road network updating based on floating car data. Geosci. Remote Sens. Symp. 2011, 24, 1878–1881.

Zheng, L.; Sun, M.; Luo, Y.; Song, X.; Yang, C.; Hu, F.; Yu, M. Utilizing MapReduce to Improve Probe-Car Track Data Mining. ISPRS Int. J. Geo-Inf. 2018, 7, 287. [CrossRef]

Zheng, Z.; Luo, C.; Ye, W.; Ning, J. Spark-Based Iterative Spatial Overlay Analysis Method. In Proceedings of the 2017 International Conference on Electronic Industry and Automation (EIA 2017), Suzhou, China, 23–25 June 2017.

Zhong, S.; Di, L.; Sun, Z.; Xu, Z.; Guo, L. Investigating the Long-Term Spatial and Temporal Characteristics of Vegetative Drought in the Contiguous United States. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2019, 12, 836–848. [CrossRef]

Zhong, S.; Fang, Z.; Zhu, M.; Huang, Q. A geo-ontology-based approach to decision-making in emergency management of meteorological disasters. Nat. Hazards 2017, 89, 531–554. [CrossRef]

Zhong, S.; Xu, Z.; Sun, Z.; Yu, E.; Guo, L.; Di, L. Global vegetative drought trend and variability analysis from long-term remotely sensed data. In Proceedings of the 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Istanbul, Turkey, 16–19 July 2019; pp. 1–6.

Zhou, Y.; Yang, L.; Van de Walle, B.; Han, C. Classification of microblogs for support emergency responses: Case study Yushu earthquake in China. In Proceedings of the 2013 46th Hawaii International Conference on System Sciences, Wailea, Maui, HI, USA, 7–10 January 2013; pp. 1553–1562.

Zhu, Y.; Zhou, W.; Xu, Y.; Liu, J.; Tan, Y. Intelligent Learning for Knowledge Graph towards Geological Data. Sci. Program. 2017, 2017, 1–13. [CrossRef]

Zhuang, C.; Xie, Z.; Ma, K.; Guo, M.; Wu, L. A Task-Oriented Knowledge Base for Geospatial Problem-Solving. ISPRS Int. J. Geo-Inf. 2018, 7, 423. [CrossRef]

Aplikasi Geografis Dengan Komputasi Big Data

Dr. Joseph Teguh Santoso, S.Kom, M.Kom

BIODATA PENULIS

Dr. Joseph Teguh Santoso, S.Kom, M.Kom adalah Rektor dari Universitas Sains & Teknologi Komputer (Universitas STEKOM) Semarang yang memiliki banyak pengalaman praktis dalam bidang *e-commerce*sejak Tahun 2002. Beliau mempunyai 3 (tiga) toko *Official Online Store* di China untuk merek Sepeda Raleigh, dengan omzet tahunan pada Tahun 2019 mencapai lebih dari Rp. 35 Milyar rupiah dan terus meningkat. Dr. Joseph T.S memiliki lisensi tunggal sepeda merek "Raleigh" untuk penjualan *Online* di seluruh China. Di samping itu beliau juga memiliki pabrik sepeda dan sepeda listrik merek "Fengjiu", yaitu Pabrik Sepeda Listrik yang masih tergolong kecil di China. Pengalaman beliau malang melintang di dunia *online store* di China seperti Alibaba, Tmall, Taobao, JD, Aliexpress sangat membantu mahasiswa untuk memiliki pengalaman teknis dan praktis untuk membuka toko *online* bersama beliau.

YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

ISBN 978-623-8120-80-2 (PDF)
9 786238 120802

Dr. Joseph Teguh Santoso, S.Kom, M.Kom

Aplikasi Geografis Dengan Komputasi Big Data

YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id