TUGAS DATA BESAR

“Resume Artikel Ilmiah”

Disusun Oleh:

Roikhatul Miskiyah - 24060119120021

Dosen Pengampu:

Satriyo Adhy, S.Si., M.T.

DEPARTEMEN ILMU KOMPUTER/ INFORMATIKA

FAKULTAS SAINS DAN MATEMATIKA

UNIVERSITAS DIPONEGORO

SEMARANG

2022

RESUME ARTIKEL ILMIAH

"Using Network Analysis and Machine Learning to Identify Virus Spread Trends in COVID-19"

ABSTRAK

Wabah Coronavirus Disease 2019 (COVID-19) telah menginfeksi dan membunuh jutaan orang di seluruh dunia, mengakibatkan pandemi dengan dampak global yang sangat besar. Penyakit ini mempengaruhi sistem pernapasan dan agen virus yang menyebabkannya adalah SARS-CoV-2 yang menyebar melalui tetesan air liur, serta melalui batuk dan bersin. Sebagai infeksi virus yang sangat menular, COVID-19 menyebabkan kerusakan signifikan pada ekonomi negara maju dan negara berpenghasilan rendah dan menengah karena dampak langsungnya tentang kesehatan warga dan langkah-langkah penahanan yang diambil untuk mengurangi virus. Metode untuk mengurangi atau mengendalikan penyebaran virus dan melindungi populasi global diperlukan untuk menghindari kematian lebih lanjut, masalah kesehatan jangka panjang, dan dampak ekonomi yang berkepanjangan. Pendekatan paling efektif untuk mengurangi penyebaran virus dan menghindari keruntuhan substansial dari sistem kesehatan, dengan tidak adanya vaksin, adalah *Nonpharmaceutical Interventions* (NPI) seperti menegakkan pembatasan penahanan sosial, memantau populasi secara keseluruhan mobilitas, menerapkan tes virus secara luas, dan meningkatkan langkah-langkah kebersihan. Pendekatan pada penelitian ini terdiri dari penggabungan analitik jaringan dengan model pembelajaran mesin dengan menggunakan kombinasi anonim data kesehatan dan telekomunikasi untuk lebih memahami korelasi antara perpindahan penduduk dan penyebaran virus. Pendekatan ini, yang disebut analisis jaringan lokasi/ *Location Network Analysis* (LNA), memberikan kemungkinan prediksi yang akurat dari kemungkinan wabah baru. Ini memberi pemerintah dan otoritas kesehatan alat penting yang dapat membantu menentukan metrik kesehatan masyarakat yang lebih akurat dan dapat digunakan untuk mengintensifkan kebijakan penahanan sosial untuk menghindari penyebaran lebih lanjut atau untuk memudahkan mereka membuka kembali perekonomian. LNA juga dapat membantu untuk mengevaluasi secara retrospektif efektivitas respon kebijakan terhadap COVID-19 (Reis Pinheiro et al., 2021).

1. Pengenalan

Pada Januari 2020, virus yang sebelumnya tidak dikenal diidentifikasi dan kemudian diberi nama 2019 novel coronavirus. Pada Februari 2020, novel coronavirus ini diberi nama Coronavirus Disease 2019 oleh Organisasi Kesehatan Dunia. Virus penyebab COVID-19 dikenal sebagai SARS-CoV-2. Pada Maret 2020, WHO menyatakan penyakit coronavirus 2019 sebagai pandemi global.

Pandemi didefinisikan sebagai penyakit yang terjadi di wilayah geografis yang luas dan menginfeksi sebagian besar populasi. Penelitian ini bertujuan untuk mengkorelasikan perilaku mobilitas, atau bagaimana orang berpindah antar wilayah geografis, dengan penyebaran virus, dan kemudian memprioritaskan lokasi geografis untuk analisis epidemiologi lebih dalam dan identifikasi populasi berisiko. Pendekatan umum penegakan

kebijakan selama pandemi penyakit menular sebagian besar bersifat reaktif. Pejabat kesehatan masyarakat melacak perubahan dalam kasus aktif, mengidentifikasi *hot spot* dengan jumlah kasus positif yang ditemukan, dan menegakkan kebijakan penahanan terutama berdasarkan kedekatan geografis.

Studi ini mengusulkan pendekatan proaktif, berdasarkan perilaku mobilitas di wilayah geografis dan korelasinya dengan penyebaran virus dari waktu ke waktu. Pergerakan populasi di seluruh wilayah geografis dievaluasi dari waktu ke waktu, dan serangkaian analisis korelasi dilakukan untuk mengidentifikasi lokasi yang memainkan peran kunci dalam arus orang dan bagaimana arus ini memengaruhi cara virus menyebar dari waktu ke waktu dan di lokasi yang berbeda. Di masa lalu, beberapa peneliti dan praktisi telah menggabungkan analisis jaringan dengan studi epidemiologi untuk memodelkan mobilitas dan memprediksi penyebaran penyakit pada manusia, diantaranya adalah Pierson dkk mengenai jaringan mobilitas terintegrasi dengan model yang rentan, terpapar, menular, dan pulih (SEIR) agar sesuai degan lintasan infeksi. Selanjutnya terdapat penelitian dari Morris yang berfokus pada pengumpulan dan analisis data dengan menggunakan formulasi jaringan penyakit menular.

Studi ini menggunakan kombinasi data telekomunikasi dan kesehatan masyarakat. Data ini diubah menjadi informasi mobilitas. Dalam penelitian ini, data kesehatan masyarakat berupa jumlah kasus baru positif COVID-19, per waktu, dan per lokasi geografis. Data mobilitas dikumpulkan setiap hari di tingkat kotamadya agar sesuai dengan perincian data kesehatan masyarakat yang tersedia.

Serangkaian analitik jaringan dilakukan pada data gabungan sehingga penelitian ini dapat lebih memahami perilaku pergerakan populasi dan korelasinya dengan penyebaran virus dari waktu ke waktu dan berdasarkan geografi. Memahami perilaku mobilitas ini dan bagaimana korelasinya dengan penyebaran virus memungkinkan otoritas kesehatan masyarakat untuk membuat keputusan yang lebih proaktif tentang kebijakan penahanan sosial.

2. Data dan Metodologi

2.1. Deskripsi Data

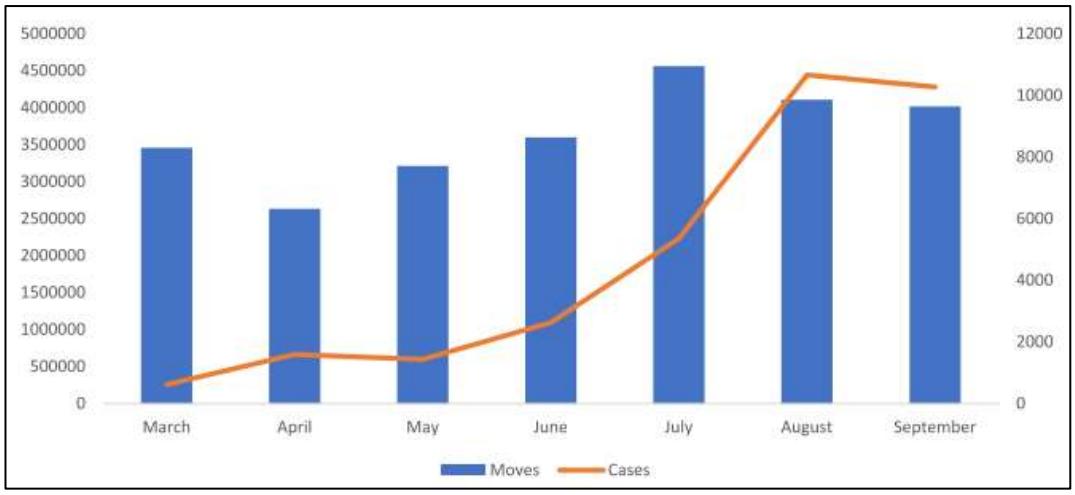
Data mobilitas yang disediakan oleh perusahaan telekomunikasi dikumpulkan oleh kotamadya setiap hari untuk dicocokkan dengan kasus positif yang disediakan oleh DOH (*The Department of Health*). Studi ini mempertimbangkan 1.551 kota dan tujuh bulan data mobilitas. 1 menunjukkan jumlah pergerakan per bulan, bervariasi

dari 3,4 hingga 4,6 juta pergerakan, dan jumlah kasus positif COVID-19, bervariasi dari 618 hingga 10.659 kejadian.

2.2. Metodologi

Location Network Analysis (LNA) berfokus pada analisis jaringan dengan menggunakan data georeferensi dari waktu ke waktu. Metode ini mengevaluasi pergerakan orang dari waktu ke waktu di wilayah geografis yang berbeda. Studi ini menerapkan LNA untuk mengkorelasikan pergerakan populasi secara keseluruhan dengan penyebaran virus corona dari waktu ke waktu dengan mempertimbangkan beberapa wilayah geografis. Data telekomunikasi yang digunakan dalam penelitian ini disediakan oleh operator telekomunikasi utama di Filipina dan data kesehatan disediakan oleh otoritas kesehatan Filipina setempat. Data dianonimkan dan dikumpulkan. Tidak ada informasi pelanggan individu atau informasi kasus positif COVID-19 individu dalam data yang dianalisis dalam penelitian ini. Tujuan utama dari studi ini adalah untuk memprediksi dan mengidentifikasi area geografis tertentu yang menjadi target kebijakan penahanan sosial, baik untuk mendefinisikan tindakan perlindungan di tempat dengan lebih baik atau untuk secara bertahap mengevaluasi lokasi mana yang siap untuk tindakan penahanan yang akan dilonggarkan.

Skenario umum analisis ini adalah spasiotemporal. Dalam analisis spatiotemporal, geolokasi dapat memiliki tingkat perincian yang berbeda misalnya, koordinat tertentu, poligon, lingkungan, wilayah administratif, dll. Dalam penelitian ini, setiap geolokasi diwakili oleh garis lintang dan garis bujur dari salah satu dari 1.551 kotamadya di Filipina. Gambar 1 merupakan jumlah grafik batang perpindahan dan kasus positif COVID-19 per bulan.



Gambar 1. Jumlah Perpindahan dan Kasus Positif COVID-19 Per Bulan

2.2.1. Esktraksi Topologi

Ekstraksi topologi adalah teknik standar dalam analisis jaringan. Dua metode spesifik yang menunjukkan korelasi tinggi dengan penyebaran virus dari waktu ke waktu adalah deteksi komunitas dan dekomposisi k-core. Dalam penelitian ini menggunakan algoritma Louvain, yang mempartisi node menjadi komunitas dengan mengoptimalkan modularitas secara heuristic. K-core memiliki banyak aplikasi praktis misalnya, mendeskripsikan jejaring sosial, memvisualisasikan grafik kompleks, menentukan peran dalam jaringan protein biologis, dan mempelajari penyebaran virus dalam epidemiologi

2.2.2. Metrik Sentralitas

Metrik sentralitas jaringan membantu menentukan peringkat lokasi menurut lalu lintas mobilitasnya. Dalam penelitian ini, sentralitas pengaruh orde pertama dan kedua menyajikan korelasi tertinggi dengan penyebaran virus. Lokasi dengan sentralitas pengaruh tinggi adalah lokasi yang memiliki kemungkinan lebih tinggi untuk menyebarkan virus ke berbagai wilayah geografis.

2.2.3. Korelasi Pearson

Korelasi *product-moment Pearson* adalah ukuran parametrik dari hubungan linier antara dua variabel. Menentukan himpunan Ω sebagai metrik jaringan berikut: komunitas, k-core, derajat, pengaruh, kedekatan, antara, koefisien pengelompokan lokal, hub, otoritas, dan PageRank. Untuk setiap metrik jaringan $C \epsilon \Omega$, di setiap geolokasi g, dan jangka waktu t, penelitian ini menghitung dua koefisien korelasi Pearson.

3. Hasil Diskusi

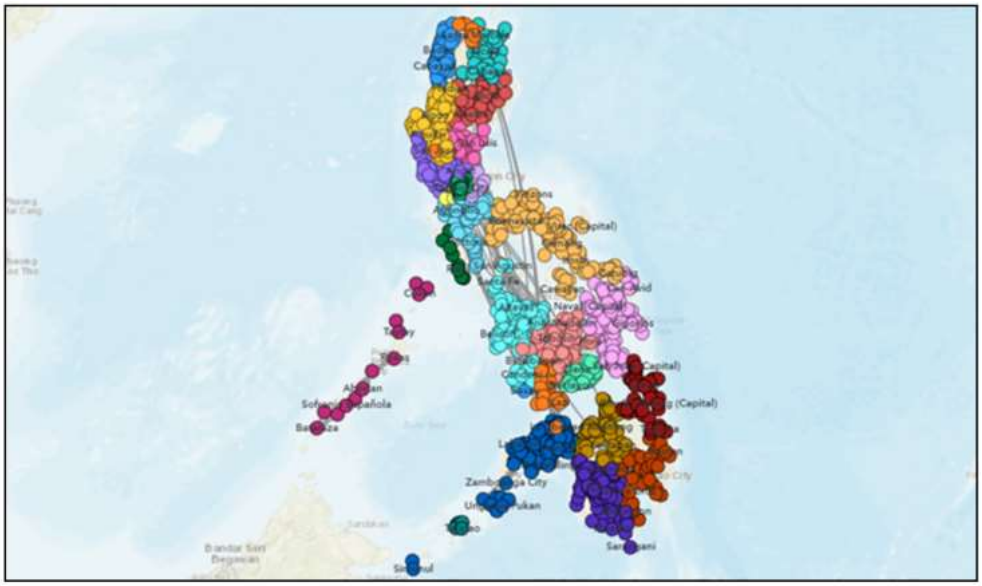
3.1. Ekstraksi Topologi

Dalam penelitian ini, node adalah geolokasi dan hubungan di antara mereka ditimbang oleh pergerakan populasi. Di bagian berikut, kita melihat komunitas dan inti untuk lebih memahami bagaimana pola topologi dalam jaringan berhubungan dengan potensi penyebaran virus.

3.1.1. Deteksi Komunitas

Algoritme deteksi komunitas mengelompokkan lokasi menurut kepadatan volume pergerakan di antara mereka. Gambar 2 menunjukkan

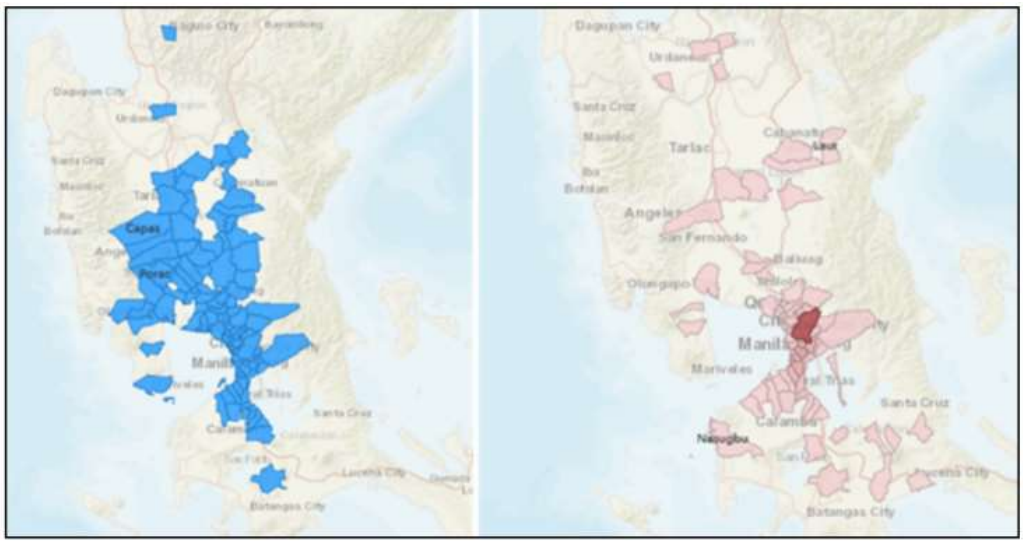
pengelompokan lokasi yang ditentukan oleh deteksi komunitas berdasarkan jumlah orang yang mengalir di antara mereka. Sebagian besar komunitas yang diidentifikasi secara geografis berdekatan satu sama lain. Hasil ini mungkin menunjukkan bahwa kebanyakan orang cenderung melakukan perjalanan ke lokasi terdekat atau pada akhirnya mereka harus melewati lokasi lain untuk mencapai tujuan akhir mereka.



Gambar 2. Komunitas terdiri dari lokasi-lokasi yang berdekatan secara geografis. Ketika virus menyerang satu lokasi, lokasi lain dalam komunitas yang sama berisiko lebih tinggi.

3.1.2. Dekomposisi K-core

Menggunakan dekomposisi k-core, penelitian ini mengelompokkan lokasi berdasarkan tingkat interkonektivitas yang serupa di seluruh wilayah. Penelitian ini tertarik pada inti yang paling kohesif dalam jaringan. Inti yang paling kohesif menghadirkan tingkat interkonektivitas yang tinggi antara lokasi di dalam teras. Salah satu hasil terpenting dari dekomposisi k-core adalah korelasi yang tinggi dengan penyebaran virus yang lebih luas. Ketika inti diidentifikasi, khususnya yang lebih kohesif, kebijakan penahanan sosial dapat dibuat lebih proaktif dalam mengidentifikasi kelompok lokasi yang harus dikarantina bersama, daripada hanya didasarkan pada kedekatan geografis dengan titik panas saat ini. Ini menjelaskan penyebaran virus dari waktu ke waktu di seluruh lokasi yang secara geografis jauh dari satu sama lain tetapi dekat dalam hal interkonektivitas, seperti yang ditunjukkan pada Gambar 3.



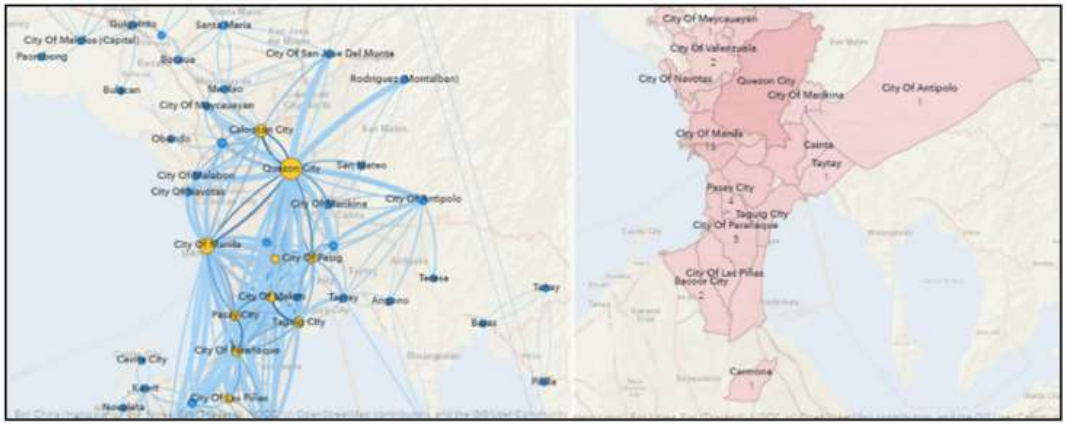
Gambar 3. Dekomposisi inti mengelompokkan lokasi berdasarkan tingkat interkonektivitas (aliran orang) di antara mereka. Peta berdampingan menunjukkan korelasi antara penyebaran virus dan inti yang paling kohesif. Di sebelah kiri, lokasi dengan warna biru tua berada di inti yang paling kohesif. Di sebelah kanan, lokasi-lokasi ini adalah tempat-tempat di mana jumlah kasus baru positif COVID-19 (berwarna merah) meningkat.

3.2. Metrik Sentralitas

Metrik sentralitas dapat memainkan peran kunci dalam menjelaskan penyebaran virus. Misalnya, sentralitas kedekatan dapat mengidentifikasi lokasi utama menurut arus orang, berkontribusi paling besar terhadap kecepatan penyebaran virus di lokasi geografis yang berbeda. Sentralitas antara dapat mengidentifikasi lokasi yang berfungsi sebagai gatekeeper, yaitu lokasi yang tidak serta merta memiliki jumlah kasus positif yang tinggi tetapi berfungsi sebagai jembatan yang menghubungkan beberapa lokasi geografis, yang menyebabkan penyebaran virus lebih luas. Untuk mengeksplorasi ini lebih lanjut, penelitian ini membuat serangkaian jaringan, berdasarkan pergerakan populasi antara wilayah geografis dan informasi kesehatan tentang kasus positif di wilayah ini dari waktu ke waktu. Kumpulan sentralitas dan topologi jaringan ini dihitung untuk setiap jaringan individu, yang mewakili satu hari analisis dalam kerangka waktu penelitian.

Dalam Gambar 4, peta di sebelah kiri menyoroti (berwarna kuning) lokasi dengan metrik jaringan gabungan yang sangat tinggi W pada beberapa spesifik waktu istirahat t. Panah antara lokasi mewakili volume besar pergerakan penduduk. Dalam hal ini, volume yang lebih besar mewakili sistem transportasi umum utama, dari daerah pusat negara ke selatan. Sistem transportasi umum ini merupakan jalur komuter penting antara wilayah geografis yang berbeda di dalam negeri, memindahkan sejumlah besar orang dari waktu ke waktu. Jumlah orang yang

bepergian melintasi lokasi geografis tersebut meningkatkan kemungkinan penyebaran virus ke wilayah yang lebih luas, terutama ke lokasi di sepanjang jalur transportasi. Seiring berjalannya waktu, peningkatan jumlah kasus positif dapat dilihat pada peta di sebelah kanan (warna merah gelap) di sepanjang wilayah yang terhubung dengan lokasi utama (berwarna kuning pada peta di sebelah kiri).



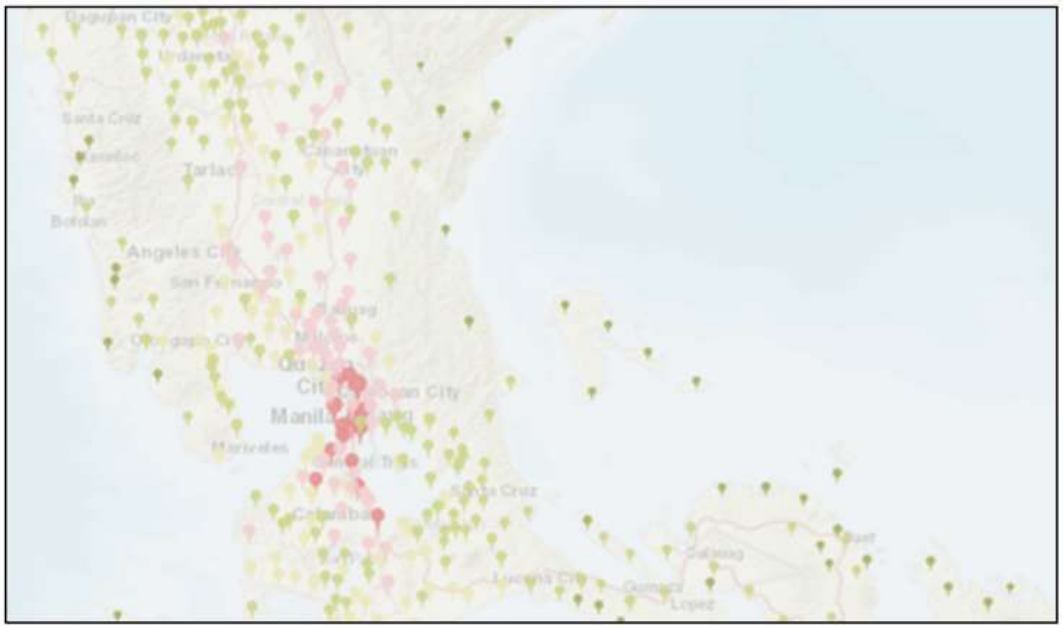
Gambar 4. Metrik jaringan menjelaskan korelasi antara perilaku mobilitas dan penyebaran virus.

3.3. Menghitung Tingkat Resiko

Dalam serupa cara, kita dapat menggunakan korelasi (μ.C gt)antara jaringan bertemu- { rics dan jumlah kasus positif di lokasi untuk lokasi asalGuntuk mengkategorikan risiko, sebagai berikut:

$$R_{gt} = \sum_{C \in \Omega} \mu_{gt}^{C} z_{gt}^{C}$$

Tingkat risiko dikelompokkan menjadi lima kelompok, menggunakan *algoritme binning* standar. Bin pertama memiliki sekitar 1% dari lokasi; itu dianggap berisiko tinggi. Semua lokasi di bin pertama ini memiliki risiko tertinggi untuk peningkatan jumlah kasus positif dari waktu ke waktu karena koneksi yang masuk. Bin kedua memiliki sekitar 3% –4% dari lokasi dan dianggap risiko sedang-tinggi. Bin ketiga memiliki sekitar 5% dari lokasi dan dianggap risiko sedang. Bin keempat memiliki sekitar 40% dari lokasi dan dianggap risiko sedang-rendah. Terakhir, bin kelima memiliki sekitar 50% dari lokasi dan dianggap resiko rendah. Semua kelompok ditunjukkan pada peta di Gambar 5.



Gambar 5. Warna hijau mewakili lokasi dengan risiko infeksi rendah, dan warna merah mewakili lokasi dengan risiko infeksi tinggi.

3.4. Menggunakan Model *Machine Learning* Untuk Memprediksi Wabah Baru

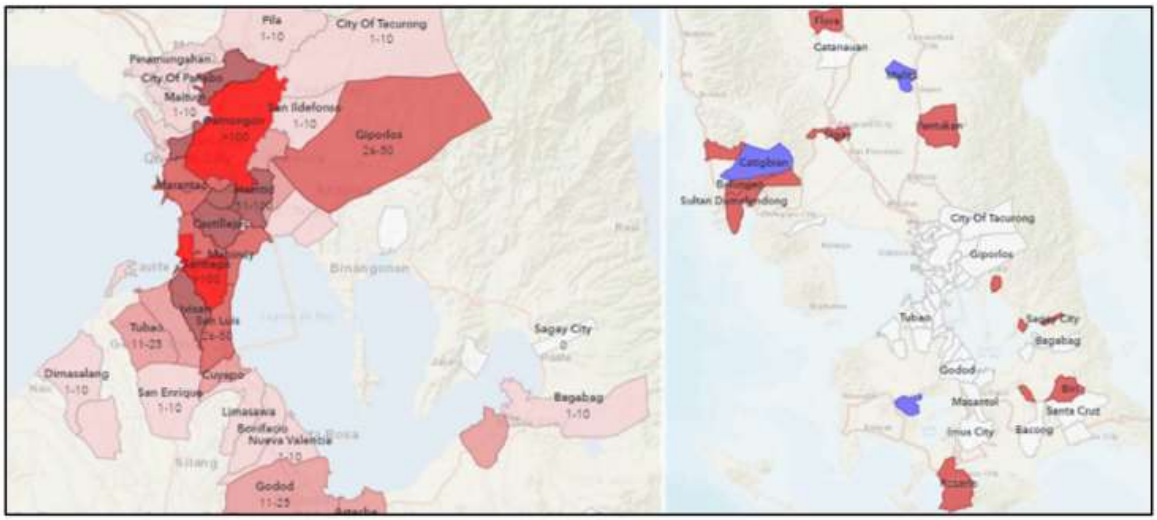
Metrik jaringan yang dipilih menghubungkan perilaku mobilitas dengan penyebaran virus dari waktu ke waktu. Karena korelasi yang jelas ini, penelitian ini telah menggunakan rangkaian ukuran jaringan ini sebagai prediktor (fitur) dalam beberapa model klasifikasi pembelajaran mesin yang diawasi. Selain variabel asli (pusat jaringan dan topologi yang diekstraksi), penelitian ini telah menghitung satu set variabel turunan baru untuk lebih menggambarkan bagaimana jaringan berkembang dari waktu ke waktu dan mempengaruhi jumlah kasus positif di beberapa lokasi. Sebagian besar variabel turunan didasarkan pada rasio metrik jaringan dari waktu ke waktu. Misalnya, untuk beberapa metrik jaringan, penelitian ini mempertimbangkan hari pertama dan terakhir dalam seminggu, hari terakhir dalam seminggu, dan nilai rata-rata untuk minggu tersebut dengan menggunakan rasio nilai maksimum dan minimum, rentang, dan standar deviasi.

Kumpulan model pembelajaran mesin terawasi meliputi regresi logistik, pohon keputusan, hutan acak, peningkatan gradien, jaringan saraf, dan model mesin vektor pendukung.

$$T = \begin{cases} 1, & \text{if } c_{gw} > c_{gw-1} \\ 0, & \text{otherwise} \end{cases}$$

4. Kesimpulan

Analisis *spatiotemporal* perilaku mobilitas dapat mengungkapkan tren penting tentang penyebaran virus di seluruh wilayah geografis dari waktu ke waktu. Analisis jaringan lokasi memungkinkan otoritas kesehatan untuk memahami dampak pergerakan populasi terhadap penyebaran virus dan pada akhirnya untuk memprediksi kemungkinan wabah baru di lokasi geografis tertentu. Korelasi antara perilaku mobilitas dan penyebaran virus memungkinkan otoritas lokal untuk mengidentifikasi kelompok lokasi untuk dimasukkan ke dalam penahanan sosial, serta tingkat keparahan yang diperlukan. Analisis jaringan lokasi memberikan informasi yang akurat untuk instansi pemerintah tentang pola penyebaran virus dan jalur umum, memungkinkan mereka untuk membuat keputusan yang baik tentang penerapan kebijakan tempat penampungan, perencanaan layanan transportasi umum, dan yang paling penting, mengalokasikan sumber daya medis ke lokasi di mana perilaku mobilitas menunjukkan peningkatan substansial dalam jumlah kasus positif. Analisis perilaku mobilitas juga dapat digunakan untuk mengidentifikasi lokasi geografis dengan risiko penularan yang lebih rendah sehingga pihak berwenang dapat mulai melonggarkan pembatasan jarak sosial dan memungkinkan aktivitas ekonomi dilanjutkan.



Gambar 6. Di sebelah kiri, lokasi yang mengalami kasus positif dalam minggu ini. Di sebelah kanan, lokasi diprediksi memiliki kasus baru pada minggu berikutnya. Merah tua mewakili kasus positif sejati, lokasi yang diklasifikasikan dengan benar oleh model. Ungu mewakili kasus positif palsu, lokasi yang salah diklasifikasikan memiliki kasus positif baru pada minggu berikutnya.

Studi ini menunjukkan bahwa otoritas kesehatan dapat menggunakan hasil LNA untuk mengendalikan penyebaran virus secara signifikan dan proaktif, karena lokasi dipantau secara ketat dengan menggunakan data geoposisi telekomunikasi anonim dan informasi infeksi dari waktu ke waktu. Teknik-teknik ini juga dapat digunakan untuk mempelajari langkah-langkah pemerintah secara surut dan mengevaluasi dampaknya

terhadap pergerakan populasi dan penyebaran virus. Memantau tingkat pergerakan penduduk dari waktu ke waktu memungkinkan otoritas lokal menentukan tindakan untuk mengendalikan penyebaran virus dengan lebih baik, mengidentifikasi tingkat risiko dalam membuka atau menutup lokasi geografis tertentu di seluruh negeri.

Analisis perilaku mobilitas dapat digunakan untuk semua jenis penyakit menular, mengevaluasi bagaimana pergerakan populasi dari waktu ke waktu dapat memengaruhi penyebaran virus di lokasi geografis yang berbeda. Metodologi ini dapat menjadi alat penting bagi kesehatan dan otoritas lokal untuk digunakan dalam menentukan tindakan sosial yang lebih akurat untuk menahan penyebaran penyakit menular dan secara global mengurangi tingkat penularan di antara berbagai wilayah.

DAFTAR PUSTAKA

Reis Pinheiro, C. A., Galati, M., Summerville, N., & Lambrecht, M. (2021). Using Network Analysis and Machine Learning to Identify Virus Spread Trends in COVID-19. *Big Data Research, 25*. https://doi.org/10.1016/j.bdr.2021.100242