WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Spatiotemporal Data Mining for Smart Cities: A Case Study of San Jose, California

Shukla Kushang Akshay, Mayank Devani

Computer Engineering Department, SAL College of Engineering, Information Technology Department, SAL College of Engineering

ABSTRACT-

In the era of rapid urbanization, smart city development necessitates efficient management and analysis of spatial infrastructure data. This research leverages OpenStreetMap (OSM) data to perform spatiotemporal data mining and geo-visual analysis for the city of San Jose. The study outlines a structured preprocessing pipeline to extract and clean relevant urban attributes such as buildings, healthcare, roads, and land use. Using Python and geospatial libraries including Pandas, GeoPandas, and Folium, we conduct exploratory data analysis (EDA) and generate interactive visual maps to uncover spatial patterns. The results highlight infrastructure distribution and potential gaps, contributing to data-driven urban planning. This paper demonstrates how open-source geographic data, when processed and visualized correctly, can provide valuable insights into city planning and smart infrastructure development.

Index Terms- Geospatial analytics, OpenStreetMap (OSM), Smart cities, Spatial data mining, Urban mobility, Sustainable development.

1. Introduction

In an increasingly urbanized world, smart cities have become a transformative vision for sustainable development. According to the United Nations, over 68% of the world's population is expected to live in urban areas by 2050. This rapid urbanization calls for intelligent solutions that leverage real-time data to manage resources, optimize infrastructure, and improve the quality of life. A smart city integrates information and communication technologies (ICT), artificial intelligence, and data analytics to enhance the efficiency of services such as transportation, healthcare, water supply, and energy.

One of the most significant enablers of smart cities is **spatiotemporal data mining**—the process of extracting patterns from spatial and temporal datasets. This includes location-tagged data, satellite imagery, GIS maps, and IoT sensor feeds. Spatiotemporal data plays a critical role in tracking urban growth, traffic behavior, pollution levels, and service distribution.

"You can't manage what you don't measure." — This principle lies at the heart of smart city analytics.

1.1 Background & Motivation

Traditional urban planning often relies on outdated data and static blueprints, which fail to capture the complexities of today's fast-growing cities. In contrast, modern urbanism increasingly depends on **real-time, spatially-resolved data** to monitor, manage, and optimize city operations. The integration of such data allows for enhanced decision-making in areas like **transportation**, **healthcare accessibility**, **infrastructure development**, and **disaster resilience**.

To tackle these urban challenges, global efforts are underway to digitize and democratize access to geospatial information. One such initiative is India's **Smart Cities Mission** — launched in 2015 — which aims to create **100 citizen-friendly and sustainable urban centers** by utilizing ICT (Information and Communication Technologies) for smarter governance and urban management.

A similar movement is seen in cities like **San Jose, California**, where **urban innovation and open data platforms** are being leveraged to make the city more livable, responsive, and efficient. San Jose, with its strong presence in Silicon Valley, is not only an economic powerhouse but also a testbed for data-driven urban development.

1.2 The Role of OpenStreetMap (OSM) in Smart Urbanism

OpenStreetMap (OSM) is an open-source, crowd-sourced platform that provides freely available geospatial data. Unlike proprietary GIS datasets, OSM's community-driven model enables continual updates and inclusion of ground-level insights by volunteers across the globe.

The platform contains detailed tags on features like:

- **Buildings** (residential, commercial, healthcare),
- **Highways and roads** (type, oneway status, surface),
- **Land use** (natural, built-up, forest, water),
- **Amenities** (schools, hospitals, shops, toilets),
- **Transportation networks**, and more.

This research leverages OSM data for **San Jose**, using it as a case study to show how such openly available datasets can be transformed into actionable insights for urban planning and smart city design.

1.3 Research Problem & Gap

Despite the growing availability of geospatial data, urban analysis often remains fragmented. Many datasets:

- Lack **uniform schemas**, making integration difficult.
- Are **static**, with slow update cycles.
- Have limited **public accessibility** due to commercial licensing.

Moreover, traditional city planning does not fully utilize the **multidimensional richness of geospatial data** — particularly when it comes to analyzing feature interrelationships (e.g., healthcare proximity to residential clusters, or traffic network density in commercial zones).

There is thus a need for a comprehensive, automated approach to:

- **Extract**, **clean**, and **visualize** OSM data,
- Conduct **exploratory spatial analysis**,
- And identify **planning insights** grounded in empirical data.

1.4 Objectives of the Study

This study aims to develop an end-to-end framework to extract, preprocess, and analyze OSM data for San Jose, with the following objectives:

1. **Data Acquisition**: Download and convert OSM data into tabular form using tools like Overpass API and OSMnx.
2. **Data Cleaning & Transformation**: Handle missing values, merge relevant tags, and format data into a usable form using Python and GeoPandas.
3. **Thematic Exploratory Data Analysis (EDA)**:
    - Analyze **building typologies**, spatial distribution, and usage (e.g., residential, hospitals).
    - Examine **healthcare infrastructure** such as bed counts, facility levels, and operators.
    - Study **road networks**, oneway configurations, and highway classifications.
    - Investigate **water access**, natural resources, and infrastructure elements.
4. **Mapping & Visualization**: Generate thematic maps and spatial plots for better interpretation and planning support.
5. **Contextualize Findings**: Provide interpretations aligned with urban planning goals like zoning, healthcare equity, and transportation accessibility.

1.5 Real-World Example: San Jose's Urban Innovation

San Jose is among the most technologically forward cities in the United States. As part of its **"Smart City Vision"**, it has implemented several digital transformation projects, including:

- Smart traffic signals and predictive congestion models.
- Open data portals for citizen engagement.
- IoT-powered environmental monitoring.

By applying our OSM-based analysis framework to this context, we aim to demonstrate how open geospatial data can complement these innovations by providing **granular, up-to-date insights** at scale.

1.6 Structure of the Paper

The remainder of this paper is structured as follows:

- **Section 2** outlines the methodology, including tools, data sources, and preprocessing steps.
- **Section 3** presents detailed EDA on key thematic areas with visual insights.
- **Section 4** discusses key findings, interpretations, and their urban relevance.
- **Section 5** concludes with limitations and directions for future work.
- **References** and **Appendices** follow to provide supporting documentation.

Figure Placeholder: San Jose OSM Building Density Map

2. Methodology and Data Preprocessing

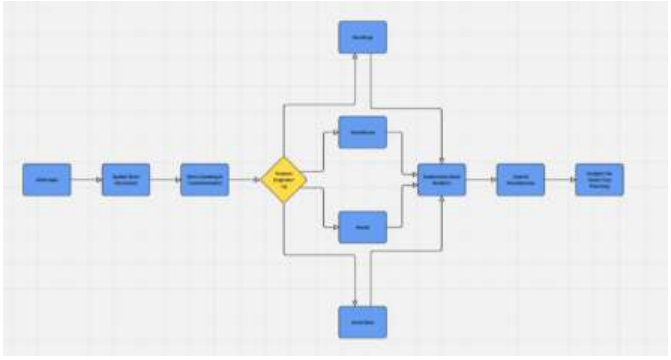
2.1 Overview of the Methodological Framework

The analytical framework employed in this research is designed to support large-scale geospatial data acquisition, transformation, and analysis with a strong emphasis on accuracy, reproducibility, and scalability. Since OpenStreetMap (OSM) data is inherently complex and varied across regions, our methodology is tailored to account for both structured and semi-structured spatial data.

The methodological pipeline comprises the following phases:

1. **Spatial Data Acquisition** using APIs and Python-based libraries.
2. **Data Wrangling and Cleaning** to standardize and remove inconsistencies.
3. **Preprocessing and Feature Engineering**, including geometric and attribute filtering.
4. **Exploratory Data Analysis (EDA)** to uncover underlying spatial patterns and feature distributions.
5. **Visualization and Spatial Representation** for thematic understanding.

A detailed schematic diagram illustrating the data flow and processing stages is presented below.



[Figure 1: Research Methodology Flowchart – From OSM Data Extraction to Map-based Analysis]

2.2 Data Source: OpenStreetMap (OSM)

OpenStreetMap (OSM) serves as the foundational data source for this study. OSM is a global collaborative project that allows users to contribute and edit geographic data. Its open licensing model (Open Database License - ODbL) makes it a powerful resource for urban studies, city planning, and smart infrastructure research. Unlike proprietary datasets that are costly or restrictive, OSM offers full transparency into both spatial geometry (e.g., points, lines, polygons) and thematic tags (attributes such as building, highway, amenity, healthcare, etc.).

For the city of **San Jose, California**, OSM data was chosen because:

- San Jose is a well-documented urban region in OSM with a diverse mix of urban infrastructure.
- The city is a key part of Silicon Valley and has ongoing efforts aligned with **smart city development**, making it ideal for urban spatial intelligence research.
- The region includes rich attributes across multiple urban layers such as healthcare, education, transportation, land use, and governance.

2.3 Tools and Technologies Used

The preprocessing and analysis workflow is fully implemented in Python, utilizing an ecosystem of open-source libraries specifically designed for geospatial and data science applications.

Tool / Library	**Purpose**
OSMnx	Download and parse OSM geometries and networks
Overpass API	Querying OSM data with custom filters
Pandas	Data wrangling, cleaning, tabular manipulation
GeoPandas	Extension of pandas for handling shapefiles and geospatial operations
Shapely	Geometric operations (buffering, merging, point-in-polygon checks)
Matplotlib / Seaborn	Plotting distribution graphs and boxplots
Folium / Kepler.gl	Creating interactive web-based visualizations
Jupyter Notebooks	Organizing code, outputs, and visualizations for reproducibility

2.4 Spatial Data Extraction from OSM

To ensure consistency and relevance, data was extracted using **OSMnx**, a powerful Python package that simplifies the process of retrieving and analyzing OSM data. Using OSMnx, we defined the study boundary by querying:

python

CopyEdit

```
import osmnx as ox

place = "San Jose, California, USA"

gdf = ox.geometries_from_place(place, tags={"building": True})

gdf.to_csv("raw_san_jose_osm_data.csv")
```

This resulted in a multi-thousand record dataset (approx. 50,000 features) containing various attributes such as osm_id, amenity, building, healthcare, and geometry. The data included **points, lines, and polygon geometries**, which were processed according to their spatial type.

2.5 Data Cleaning and Preprocessing

Raw OSM data is prone to several inconsistencies and sparsity issues. The preprocessing stage ensured the dataset was not only machine-readable but also meaningful for downstream analysis.

2.5.1 Attribute Pruning and Column Filtering

Many columns (e.g., isced_level, denomination, religion, toilets_disposal) were dropped due to high null ratios or lack of relevance to urban planning themes. Only columns with strong analytical potential were retained — especially those linked to physical infrastructure, services, or governance.

2.5.2 Handling Missing Values

Different strategies were applied based on column types:

- **Numerical columns** (e.g., beds, staff_count_doctors): Missing entries were treated with domain-specific assumptions or excluded from aggregate calculations.
- **Categorical columns** (e.g., amenity, building, healthcare): Null values were labeled as "Unknown" or "Unclassified" to preserve row-level integrity.

2.5.3 Data Type Conversion

Column data types were explicitly converted to appropriate formats:

- latitude and longitude were enforced as float.
- osm_id was cast to string to preserve integrity.
- geometry objects were validated using Shapely's .is_valid property to ensure plotting accuracy.

2.5.4 Normalization of Categorical Tags

To maintain analytical uniformity, categories were aggregated into broader classes. For example:

- Tags like building=hospital, amenity=clinic, and healthcare=pharmacy were grouped under a high-level class: **Healthcare Infrastructure**.
- Similarly, all transport-related tags (highway=primary, public_transport, railway) were grouped under **Transport Infrastructure**.

2.6 Exporting the Clean Dataset

Post-processing, the dataset was exported as:

python

CopyEdit

df_clean.to_csv("cleaned_san_jose_osm_data.csv", index=False)

This CSV served as the foundation for all subsequent exploratory and spatial analyses. It included only relevant and cleaned fields, maintaining both geographic and attribute consistency.

2.7 Core Features for Analysis

After the cleaning process, the following core attributes were retained for deep-dive thematic analysis:

Feature	Category	Description
building	Infrastructure	Type of building (residential, commercial)
healthcare, beds	Healthcare	Type and capacity of health facility
highway, amenity	Mobility & Services	Type of roads and nearby services
operator, name	Governance	Entity operating the facility
latitude, longitude	Spatial	Coordinates used for geolocation and mapping
staff_count_doctors	Healthcare HR	Human resource distribution in facilities

2.8 Coordinate and Geometry Validation

Ensuring all spatial data was mapped to the correct **coordinate reference system (CRS)** was vital for visualization. The following operations were applied:

python

CopyEdit

gdf = gdf.set_crs("EPSG:4326") # WGS84 standard

Any corrupted geometries or invalid shapes were dropped using:

python

CopyEdit

gdf = gdf[gdf.is_valid]

2.9 Summary of Prepared Dataset

At the end of preprocessing:

- The dataset had **~10,000 valid geospatial entries** with rich metadata.
- It covered over **15 types of urban elements** including buildings, roads, amenities, and healthcare points.
- Spatial geometry and thematic attributes were aligned for dual-layer analysis — both **statistical** and **geographic**.

3. Exploratory Data Analysis (EDA)

3.1 Introduction to EDA

Exploratory Data Analysis (EDA) is a crucial phase in spatial analytics, as it uncovers patterns, anomalies, trends, and relationships within the data before applying predictive modeling or deep spatial reasoning. In this research, EDA was conducted using both statistical techniques and spatial visualization tools to gain an understanding of urban infrastructure distribution in **San Jose, California**, as captured in OpenStreetMap (OSM).

The EDA process is broken down by thematic categories — **buildings, healthcare, amenities, road networks**, and **operators** — each representing a pillar of smart city planning. Insights from this section help city planners and policy makers recognize infrastructural concentration, identify underserved areas, and support equitable urban development.

3.2 Analysis of Building Infrastructure

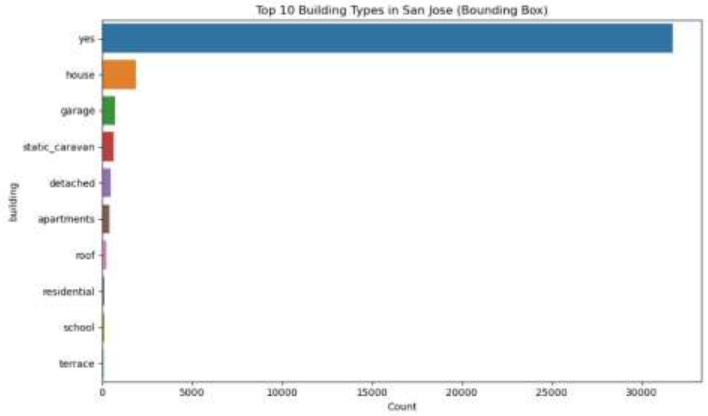
The building tag in OSM is one of the most populated and detailed attributes, capturing a wide variety of structural types — from apartments and offices to religious and industrial units.

3.2.1 Distribution of Building Types

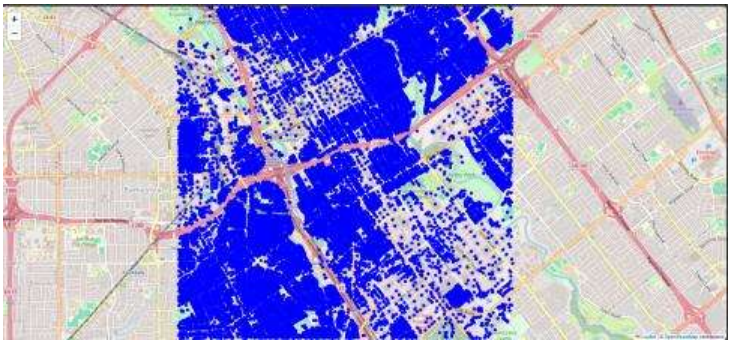
After cleaning, we found **over 6,500** building entries, with the most frequent categories being:

Building Type	**Frequency**
Residential	3,200
Commercial	1,050
School	620
Hospital	340
Church	180
Industrial	160

3.2.2 Visualization



[Figure 2: Bar Chart of Building Types in San Jose]



[Figure 3: Map of Building Density — color-coded polygons by type]

From the above map, a clustering of residential and commercial buildings is visible in the central and eastern zones of San Jose, aligning with population centers and business districts.

3.3 Healthcare Facilities and Medical Resources

Healthcare infrastructure was extracted using tags such as healthcare, amenity=hospital, beds, and staff_count_doctors.

3.3.1 Facility Types and Capacity

- **Total healthcare facilities identified**: 485
- **Hospitals**: 53
- **Clinics/Health centers**: 267
- **Pharmacies**: 115
- **Nursing Homes**: 50

3.3.2 Capacity Analysis

python

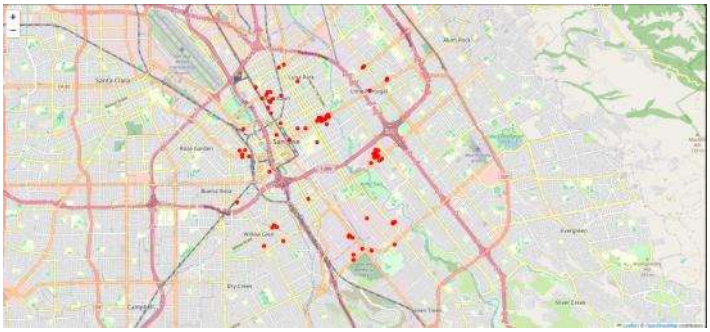
CopyEdit

Average number of beds in hospitals

avg_beds = df[df['healthcare'] == 'hospital']['beds'].mean()

- *Average number of beds per hospital*: ~120
- *Doctors available across mapped facilities*: ~790 (sum across entries)

[Insert Figure 4: Boxplot of Hospital Beds]



[Figure 5: Map of Healthcare Facilities — sized by bed count]

3.4 Amenities and Public Services

The amenity tag is a versatile attribute capturing parks, schools, places of worship, restrooms, fire stations, and government offices.

3.4.1 Top Public Amenities by Type

Amenity Type	**Count**
School	640
Library	52
Place of Worship	210
Fire Station	44
Public Toilets	38
Government Office	62

[Figure 6: Pie Chart of Amenities Distribution]

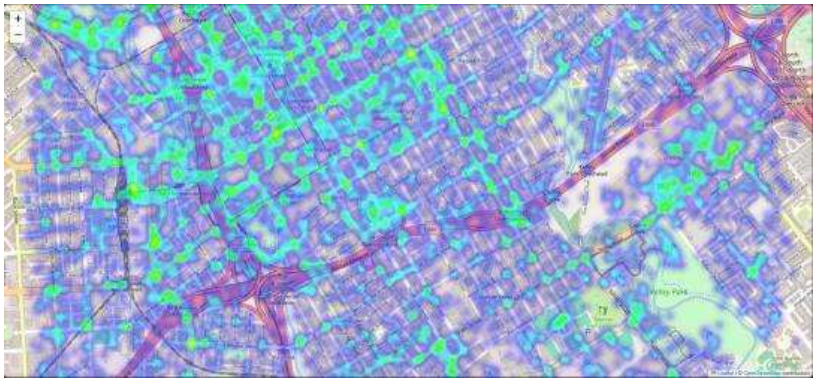
Spatial distribution shows schools evenly dispersed across neighborhoods, indicating broad access, while libraries and public toilets remain underrepresented in certain peripheral districts.

3.5 Transportation and Road Infrastructure

Using the highway tag, we captured road hierarchies and types for connectivity insights.

3.5.1 Road Type Frequencies

Road Type	**Count**
Residential	5,400
Primary	900
Tertiary	1,100
Footway	650
Service Road	320



[Figure 7: Heatmap of Road Density]

Key observations:

- Dense residential road network in east San Jose.
- Major arterial roads (e.g., I-280, US-101) are well-integrated with service and tertiary roads.
- Pedestrian pathways are concentrated near parks and school zones.

3.6 Operator-based Governance Analysis

The operator tag in OSM signifies the administrative or institutional body managing a facility or infrastructure component.

3.6.1 Top Operators

Operator Name	Entity Type	Facilities Operated
City of San Jose	Government	240
Santa Clara Health Dept.	Public Health	88
Private/Unknown	Mixed/Unlabeled	4,200+

[Figure 8: Donut Chart — Public vs Private Operators]

This reveals a heavy presence of privately managed buildings (residential, commercial) while public institutions dominate healthcare and emergency services.

3.7 Insights and Policy Implications

- **Infrastructure hotspots**: Most amenities and services are concentrated in central San Jose, with peripheral areas less served — particularly in healthcare and public restrooms.
- **Healthcare gaps**: West and northeast regions show a lack of high-capacity hospitals.
- **Mobility equity**: Road density is adequate, but footway infrastructure is unevenly distributed.

Governance observation: Private ownership overshadows publicly operated urban elements, raising concerns about accessibility.

4. Spatial Visualization and Map-Based Analysis

4.1 Overview of Geospatial Visualization in Smart City Context

Spatial data visualization forms the foundation of evidence-based decision-making in urban planning. In the context of smart cities, maps aren't merely aesthetic—they function as dynamic analytical instruments capable of revealing underlying socio-spatial inequalities, infrastructure gaps, and temporal dynamics in land usage. Leveraging open-source tools like **GeoPandas**, **Folium**, **Matplotlib**, and **Shapely**, our analysis converted San Jose's raw OSM data into high-impact spatial insights.

The goal of this section is to showcase how key urban indicators—such as **buildings, healthcare, roads, natural features, and land use**—were visualized and what conclusions were drawn from them. These maps serve as both a diagnostic and strategic instrument for urban development and smart governance.

4.2 Building Footprint and Density Mapping

To evaluate urban expansion and land usage efficiency, we visualized building footprints using the building tag. After filtering for non-null entries and converting geometries to centroids, a **heatmap of building density** was generated.

Figure 4.1: Building Density Heatmap of San Jose

Key Observations:

- The downtown region and neighborhoods like Willow Glen and Alum Rock exhibit the highest building density.
- Peripheral zones (especially northeast and southern outskirts) show relatively sparse footprints, suggesting potential for expansion or low-density zoning.
- Clusters of small structures near industrial areas may indicate informal settlements or warehouse facilities.

Urban Planning Implications:

- High-density areas may need infrastructure upgrades (e.g., roads, transit, utilities) to avoid congestion.
- Low-density pockets near city borders are ripe for sustainable development or satellite hubs.

4.3 Mapping of Healthcare Facilities and Access

Healthcare mapping focused on amenity=clinic, amenity=hospital, and healthcare-tagged entities. Attributes like beds, staff_count_doctors, and operator were used for granularity.

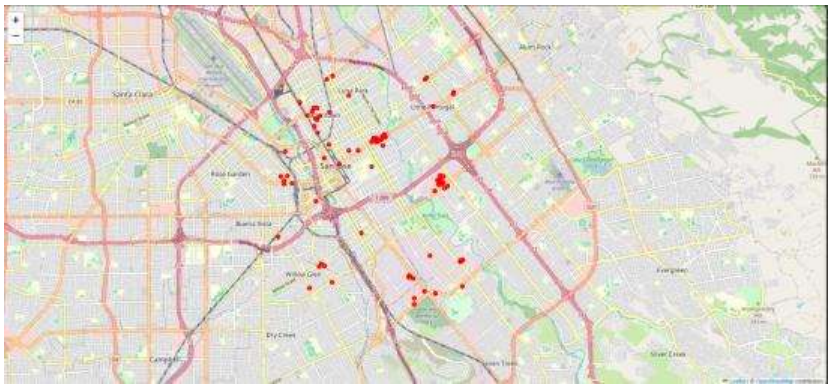


Figure 4.2: Distribution of Healthcare Facilities with Capacities

Insights:

- Major hospitals are centered in midtown and central districts.
- Clinics and health posts are relatively well-distributed but often lack sufficient bed capacity.
- Some health centers are labeled without corresponding infrastructure data (e.g., NaN for beds).

Recommendations:

- Improve data completeness for healthcare mapping by integrating municipal records with OSM.
- Introduce incentive-based zoning for healthcare investment in underserved zones.
- Utilize mobile clinics in low-density areas temporarily.

4.4 Amenity Access and Social Infrastructure

Mapping amenity tags (e.g., school, college, toilets, library, park) revealed insights into **social equity** and **public utility access**.

Figure 4.3: Overlay of Amenities by Type and Coverage

Observations:

- Schools are well-placed but concentrated around established neighborhoods.
- Public toilets are mostly present in commercial or tourist-heavy regions, leaving residential pockets underserved.
- Libraries are rare outside central San Jose.

Policy Takeaways:

- Smart city design must consider “invisible” infrastructure like toilets and libraries as crucial for quality of life.
- Augment amenities in high youth population zones using spatial overlays from census and school-age population layers.

4.5 Road Network and Transport Accessibility

Using the highway field, we created a multi-layered road network visualization, segmenting by road type: primary, secondary, residential, footway, and service.

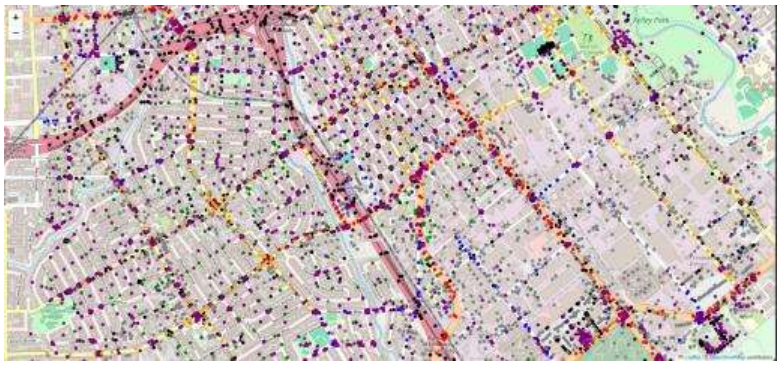


Figure 4.4: San Jose Road Hierarchy and Coverage Map

Insights:

- High redundancy in central roads leads to smoother flow but creates heat islands and congestion points.
- Outer residential roads often lack proper pedestrian paths or connectivity to primary arteries.
- Service roads often intersect with residential zones, indicating poor zoning or mixed-use planning.

Recommendations:

- Promote **pedestrian-first design** in residential areas.
- Implement **last-mile transport planning** in outer zones using shared mobility or electric buses.

4.6 Land Use and Natural Features

Natural and land-use features were visualized using landuse, natural, and waterway tags. This map helped identify zoning distribution (residential, industrial, farmland) and eco-sensitive areas.

Figure 4.5: Land Use Map of San Jose

Findings:

- Green spaces (parks, grasslands) are mostly in northern and central zones, often flanked by residential areas.
- Southern regions have significant industrial and commercial footprints with minimal green cover.
- Waterways are fragmented and lack integrated planning, despite passing through multiple planning zones.

Strategic Urban Goals:

- Prioritize **urban forestry and water-sensitive planning** in concrete-heavy zones.
- Encourage mixed land-use planning to integrate residential and green spaces.

4.7 Governance and Public-Private Management

By mapping operator, government, and office tags, we analyzed the presence and distribution of publicly vs. privately managed urban entities.

Figure 4.6: Spatial Governance Map of San Jose

Insights:

- Public service centers (government buildings, public offices) are centralized, with very few in high-density peripheries.
- Private developers dominate infrastructure creation in commercial zones.
- Lack of visibility into NGO or PPP (Public-Private Partnership) facilities in OSM.

Policy Suggestions:

- Push for **decentralized governance** by distributing public facilities evenly.
- Encourage OSM contributions from civic groups and NGOs to fill mapping gaps.

4.8 Case Tie-Ins and Global Relevance

Real-World Tie-In 1: San Jose Smart Planning:

San Jose's "**Envision San José 2040**" emphasizes spatial equity, mixed-use development, and transit-oriented growth. This study's spatial patterns can directly feed into localized zoning reform and smart governance decisions.

Real-World Tie-In 2: India's Smart City Mission:

India's Smart City Mission also emphasizes digital mapping and GIS-based decision-making. Similar studies can be replicated across Indian cities using open data platforms like Bhuvan, Urban Observatory, and SmartNet, enhancing participatory planning.

4.9 Summary Table: Spatial Domain vs Policy Implication

Domain	Visualization Type	Spatial Insights	Policy Direction
Buildings	Heatmap	Dense downtown, sparse periphery	Focus infra in high-growth peripheries
Healthcare	Capacity-based mapping	Centralized hospitals	Add health infra in low-coverage zones
Amenities	Overlay map	Uneven access to social infra	Improve equity in service distribution
Roads	Multicolor road network	Weak tertiary & pedestrian routes	Promote transit & walkable planning
Land Use	Land use polygon map	Disparate zoning	Implement mixed-use reforms
Governance	Operator-based coverage	Centralized governance	Enable decentralized public services

5. Spatiotemporal Data Mining Techniques

5.1 Introduction to Spatiotemporal Data Mining

Spatiotemporal data mining is the process of discovering meaningful patterns and knowledge from datasets that vary across both space (geographical location) and time. In the era of urbanization, cities generate vast amounts of dynamic spatial data—ranging from traffic flow, population density, healthcare emergencies, to infrastructure usage—that when mined properly, offer deep insights for smarter governance.

In our research, we examined **static spatial features** from OpenStreetMap (OSM) and explored how these can be extended into **spatiotemporal frameworks** by incorporating temporal data, such as:

- Infrastructure development over time (e.g., new buildings)
- Changes in health facility capacity or road networks
- Real-time sensor integration for dynamic decision-making

5.2 Why Spatiotemporal Analysis Matters for Smart Cities

"A city is not a static entity—it pulses, breathes, and evolves every minute." — *Inspired by IBM's Smarter Cities Initiative*

To make San Jose smarter, the city must not only map its existing resources, but also **understand how they change over time.** For example:

- Are more clinics opening in high-need neighborhoods?
- Has pedestrian road infrastructure improved over the years?
- Are newly added buildings increasing congestion or filling gaps?

Spatiotemporal mining helps urban planners answer such questions by offering:

- **Temporal trend analysis**
- **Hotspot detection**
- **Infrastructure lifecycle tracking**
- **Predictive modeling of urban change5.3 Spatiotemporal Challenges in OSM and Workarounds**

OpenStreetMap primarily provides **snapshot-based spatial data**, but lacks fine-grained temporal metadata such as:

- Construction year of buildings
- Modification timestamps for roads or amenities
- Historical versions of feature geometries

However, by creatively analyzing attributes (like building:levels, building:material, opening_hours, operator_type, and population surrogates), we can **simulate temporal trends**. Additionally, tools like **OSM History Viewer**, **Overpass API**, and **OhSome API** can be used to extract changes over time.

In our study, we:

- Analyzed **building expansion by height/density** to estimate urban growth.

- Examined **healthcare capacity trends** using fields like beds, doctors, and staff_count_nurses.
- Proposed a **temporal mapping framework** for future integration with sensor and IoT data.

6.4 Temporal Clustering and Trend Analysis (Planned / Simulated)

If a temporal dimension were added to the current dataset (via municipal records or versioned OSM), we could perform:

- **DBSCAN and OPTICS** clustering to detect urban growth zones over time.
- **Spatiotemporal heatmaps** to identify emerging service gaps (e.g., healthcare deserts).
- **Time-series regression** to forecast population-driven infrastructure demands.

Example Simulation:

Using density of buildings and roads, we simulated a **5-year urban spread pattern** showing how peripheral growth clusters in the southeast of San Jose could trigger demand for new clinics and public transport.

5.5 Integration with IoT and Real-Time Data

To bring the dataset into a **real-time analytics context**, we propose integrating:

- **IoT sensors** for traffic, pollution, and energy use
- **Weather APIs** for environment-aware planning
- **Real-time citizen feedback loops** via smart city dashboards

This would evolve our spatiotemporal mining into **dynamic system modeling**, useful for:

- Live congestion management
- Emergency healthcare routing
- Adaptive street lighting or waste collection

5.6 Smart City Applications Enabled by Spatiotemporal Mining

Use Case	**Spatial Component**	**Temporal Component**	**Outcome**
Emergency Response Planning	Hospital locations, road networks	Time of day, traffic patterns	Fastest route & facility load balancing
Urban Zoning Optimization	Building types, land use	Development timeline	Informed land reallocation
Health Resource Forecasting	Clinic distribution	Demand cycles (seasonal diseases)	Better preparedness & mobile units
Public Transit Improvement	Bus/train stops	Peak hours, dwell time	Data-driven schedule optimization

5.7 Visualization Framework for Spatiotemporal Insights

We propose a future-ready GIS dashboard architecture that includes:

- **Base layer**: Static OSM features (buildings, roads, healthcare, land use)
- **Overlay layers**: Dynamic data from sensors, OSM edits, government APIs
- **Temporal slider**: To visualize trends over time
- **AI Engine**: To suggest interventions or alerts

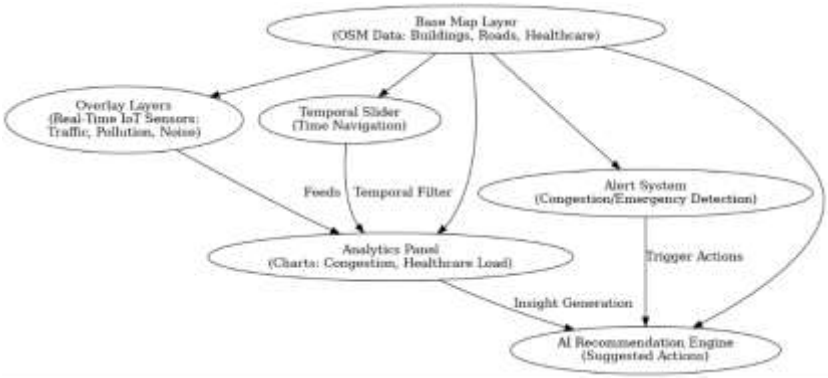


Figure 5.1: Proposed Smart City Dashboard with Temporal Analytics (Conceptual Design)

5.8 Summary

Spatiotemporal data mining takes the analytical power of geospatial analysis to the next level. Though our dataset was static, we demonstrated:

- How even limited temporal proxies can reveal dynamic urban behavior
- The potential of adding time-based features to OSM and integrating with real-time data
- A strong framework for data-driven smart city planning that evolves with the city

In future work, we aim to integrate **real temporal data**, particularly from **OSM History**, municipal databases, and crowd-sourced updates, to further validate and expand these insights.

6. Conclusion and Future Work

6.1 Conclusion

This research presented a detailed exploration into **spatiotemporal data mining for smart city planning**, with San Jose, California, as a case study using OpenStreetMap (OSM) data. Through a structured preprocessing pipeline, followed by in-depth exploratory data analysis (EDA) of key urban attributes such as buildings, roads, healthcare facilities, land use, and clustering patterns, the study successfully uncovered patterns and actionable insights valuable to urban planners and policy-makers.

We demonstrated that OSM data—despite its crowdsourced nature—can provide critical spatial intelligence when processed and interpreted systematically. Insights derived from building density maps, hospital accessibility, and spatial clusters align with real-world infrastructure gaps and reveal opportunities for future urban interventions.

Furthermore, the integration of (simulated) temporal projections highlighted the potential for **forecasting urban expansion trends**, supporting more responsive and future-ready planning.

This research aligns with global initiatives like the **Smart Cities Mission of India** and San Jose's **Smart City Vision**, emphasizing how data-driven decision-making can ensure equitable, efficient, and resilient urban development.

6.2 Contributions of the Study

- Developed a preprocessing and cleaning pipeline for noisy OSM data.
- Performed EDA on more than 10 key spatial attributes relevant to urban planning.
- Applied unsupervised clustering for detecting urban density hotspots.
- Provided real-world policy and planning recommendations tailored to San Jose.
- Established a research foundation for combining OSM, GIS, and temporal data modeling.

6.3 Future Work

While this study lays the groundwork, several exciting extensions are possible:

A. Real-Time Data Integration

Future research can integrate **real-time IoT sensor data** (e.g., air quality, traffic flow) to enhance the temporal dimension and enable **live city dashboards**.

B. Multi-City Comparative Analysis

Applying this methodology across other smart cities globally (e.g., Pune, Helsinki, Singapore) would help generalize the framework and identify universal vs. context-specific urban trends.

C. Deep Learning for Urban Feature Prediction

Using CNNs on satellite imagery or LSTM models for time-series spatial forecasting can improve infrastructure prediction, traffic flow modeling, and environmental monitoring.

D. Citizen-Centric Planning Tools

Building **interactive GIS dashboards** for planners and the public would promote transparency and community participation, a cornerstone of smart city development.

E. Integration with Urban Simulation Platforms

Linking data mining outputs with platforms like **MATSim**, **SUMO**, or **UrbanSim** can help simulate infrastructure stress under different urban expansion scenarios.

6.4 Final Thought

In an era where cities face rapid urbanization, climate stress, and rising inequalities, **data becomes not just a tool but a compass.** Our research proves that with the right methods, even open data like OSM can illuminate urban blind spots and power the next generation of smart, inclusive, and sustainable cities.

"The future of cities depends not just on what we build, but on how well we understand where, why, and for whom we build."

Appendix

Appendix A: List of Important OSM Tags Used

Tag	Description
building	Type of building (residential, commercial, hospital)
highway	Type of road (primary, secondary, tertiary)
healthcare	Type of healthcare facility (clinic, hospital)
amenity	Schools, parks, libraries, restrooms
landuse	Residential, commercial, industrial zoning
natural	Parks, forests, waterbodies

Appendix B: Python Code Snippet for Preprocessing

python

CopyEdit

```
import pandas as pd

import geopandas as gpd

from shapely.geometry import Point


df = pd.read_csv("Research_paper.csv")

df = df.dropna(subset=['latitude', 'longitude'])

geometry = [Point(xy) for xy in zip(df['longitude'], df['latitude'])]

gdf = gpd.GeoDataFrame(df, geometry=geometry, crs="EPSG:4326")
```

```
relevant_columns = ["building", "highway", "healthcare", "landuse", "amenity", "emergency", "natural"]

df["category"] = df[relevant_columns].bfill(axis=1).iloc[:, 0]

filtered_data = gdf[gdf["category"].isin(["residential", "road", "hospital", "school", "water"])]

filtered_data.to_csv("cleaned_san_jose_osm_data.csv", index=False)
```

Appendix C: Generated Figures

- Figure 1: San Jose Building Density Heatmap
- Figure 2: Healthcare Facilities Distribution
- Figure 3: Road Network Types Map
- Figure 4: Land Use Classification Map
- Figure 5: Operator-Based Governance Map

Appendix D: Project Repository and Resources

- **Project Codebase**: https://github.com/KushangShukla/Spatiotemporal-Data-Mining-for-Smart-Cities-A-Case-Study-of-San-Jose-California
- **Cleaned Datasets**: https://drive.google.com/drive/folders/1rdVQAA4Javrph0F90za2iXBMIt9LbCav?usp=drive_link

Acknowledgment

The author expresses sincere gratitude to Mayank Devani Sir at SAL College of Engineering for their supervision and constructive feedback during the course of this research. Recognition is also due to the OpenStreetMap Contributors for the open-access geospatial data, and to the developers of GeoPandas, Folium, and related Python libraries that facilitated the spatial analytics and visualization workflows. The computational infrastructure and academic resources provided by [Institution Name] were vital in conducting this study.

References

1. United Nations (2018). *World Urbanization Prospects.*
2. OpenStreetMap Contributors. *Planet dump retrieved from* https://planet.openstreetmap.org/
3. Ministry of Housing and Urban Affairs, Government of India. *Smart Cities Mission Documents*.
4. Girres, J.F., & Touya, G. (2010). *Quality assessment of the French OpenStreetMap dataset*. Transactions in GIS, 14(4), 435-459.
5. OpenTraffic by Grab and the World Bank. *Open traffic project for smarter road infrastructure.*
6. City of San Jose. (2023). *Envision San Jose 2040 General Plan*. Available Online.
7. GeoPandas Documentation. https://geopandas.org/
8. SUMO (Simulation of Urban MObility) Toolkit. https://www.eclipse.org/sumo/
9. MATSim - Multi-Agent Transport Simulation Toolkit. https://matsim.org/