GTA Real Estate Hotspots: A Graph-Based Network Approach to Predicting Next-Year Price Growth Areas

Kyle Williamson
218953901@my.yorku.ca
York University
Toronto, Ontario, Canada

Yadon Kassahun
219744291@my.yorku.ca
York University
Toronto, Ontario, Canada

Utsav Patel
219577840@my.yorku.ca
York University
Toronto, Ontario, Canada

Hari Patel
219952670@my.yorku.ca
York University
Toronto, Ontario, Canada

Abstract

We present a graph-based approach for predicting real estate hotspots in the Greater Toronto Area (GTA), modeling spatial dependencies between neighborhoods. Nodes represent Forward Sortation Areas (FSAs) and edges represent spatial connectivity. We collected 358,713 building permits from Toronto Open Data (1981-2025), aggregated to 99 FSA nodes, and constructed a spatial network with 165 connections. Network analysis reveals density of 0.034 and clustering coefficient of 0.653, indicating meaningful spatial structure. This progress report documents our data collection pipeline, network construction methodology, and initial analysis, representing approximately 50% completion toward predictive hotspot models.

1 Introduction

1.1 Motivation

The Greater Toronto Area (GTA) real estate market exhibits complex spatial and temporal dynamics, with property values influenced by accessibility, amenities, development activity, and spillover effects from neighboring areas. Traditional forecasting methods treat each location independently, ignoring fundamental spatial dependencies. Network-based methods model these relationships explicitly by representing geographic areas as nodes and their spatial connections as edges. This project evaluates whether graph-based spatial networks can effectively predict real estate hotspots and identify interpretable growth drivers using publicly available data from Toronto Open Data and OpenStreetMap.

1.2 Research Questions

This project addresses four key questions:

- Can graph-based spatial networks predict hotspots better than non-spatial baselines?
- Which features are most predictive of price growth (accessibility, amenities, development, or spatial spillover)?
- Do spatial models generalize across different GTA regions (e.g., downtown Toronto vs. outer suburbs)?
- Can we provide interpretable explanations for model predictions to support decision-making?

1.3 Team Organization

This project was completed by a four-person team with distributed responsibilities:

- **Kyle Williamson (Data Engineer)**: Data acquisition, ETL pipeline, data validation, GitHub infrastructure
- **Yadon Kassahun (Network Architect)**: Network construction, edge creation algorithms, centrality measures
- **Utsav Patel (Modeler)**: Feature engineering, baseline models, evaluation pipeline
- **Hari Patel (Analyst/Writer)**: Exploratory analysis, visualizations, documentation

The team coordinated through weekly meetings and GitHub collaboration.

1.4 Progress Report Scope

This report documents work completed through Week 3, representing approximately 50% of total project scope. We have implemented data collection infrastructure, constructed a spatial network, and conducted preliminary analysis. Sections 4.3-4.4 outline planned feature engineering and modeling for Weeks 4-6.

2 Problem Definition

2.1 Formal Problem Statement

Let $G = (V, E)$ represent a spatial network of the GTA, where:

- $V = \{v_1, v_2, \ldots, v_n\}$ are geographic cells (nodes) partitioning the GTA. Nodes represent Forward Sortation Areas (FSAs), the first three characters of Canadian postal codes.
- $E \subseteq V \times V$ are edges between spatially proximate cells. Edges $(v_i, v_j) \in E$ exist if cells i and j are within 5 km.
- Each node v_i has feature vector $\mathbf{x}_i(t)$ at time t, containing attributes such as amenity density, development activity, demographics, and historical prices.
- Historical development data $y_i(t)$ for each node, measured by building permit counts and construction values.

We define two complementary prediction tasks:

Task 1 (Regression): Predict percentage price/rent change $\Delta y_i(t+1)$ for each node v_i at the next time period.

Task 2 (Classification): Classify each node as "hotspot" (top-k growth) or "non-hotspot" based on predicted growth rates.

2.2 Success Criteria

A successful model should:

(1) Beat naive baseline in RMSE and MAE on held-out test year

(2) Show significant improvement over LASSO regression (paired t-test, $p < 0.05$)
(3) Demonstrate significant spatial autoregressive coefficient ρ in SAR model ($p < 0.05$)
(4) Achieve Precision@10 > 0.5 for identifying top growth areas
(5) Show consistent performance across Toronto core vs. outer GTA regions

2.3 Scope and Constraints

This project focuses on neighborhood-level (FSA) prediction using publicly available data, limiting scope to spatial and local factors. The network construction uses approximate geographic coordinates derived from FSA codes, sufficient for macro-level analysis but representing a limitation for fine-grained predictions.

3 Related Work

3.1 Spatial Econometrics

Spatial autoregressive (SAR) models [1, 4] extend linear regression by incorporating a spatial lag term $\rho W y$, where W is a spatial weight matrix and ρ measures spatial spillover effects. Anselin [1] demonstrated that ignoring spatial autocorrelation leads to biased estimates. Our work integrates SAR models with network-based feature engineering.

3.2 Graph Neural Networks

Graph neural networks (GNNs) enable learning on graph-structured data. Kipf and Welling [3] introduced Graph Convolutional Networks (GCNs) that aggregate features from neighboring nodes. While GCNs show promise for spatial prediction [5, 7], they require substantial data and computational resources. Our approach prioritizes interpretability and practical deployment with limited data.

3.3 Real Estate Prediction and Network Analysis

Geographically Weighted Regression (GWR) [6] allows spatially-varying regression coefficients but does not explicitly model network structure. Network analysis has been applied to urban systems for transportation [5] and infrastructure planning, but application to real estate hotspot prediction remains underexplored. Our work contributes a practical framework for constructing spatial networks from postal codes and building permits.

4 Methodology

4.1 Data Collection and Preprocessing

4.1.1 Data Sources. We collected 358,713 building permit records from Toronto Open Data spanning 1981-2025, which serve as a leading indicator of development activity and future price growth. Records include permit metadata (type, status, dates), location (address, postal code, ward), and development characteristics (structure type, use, estimated cost).

4.1.2 FSA Extraction and Aggregation. Canadian postal codes follow the format A1A 1A1, where the first three characters define the Forward Sortation Area (FSA). We extracted FSAs from the permits dataset, yielding 99 unique areas covering Toronto and surrounding municipalities. For each FSA, we aggregated: (1) total permit count, (2) construction value, and (3) temporal distribution of permits for trend analysis.

4.1.3 Coordinate Approximation. The permits dataset lacks precise coordinates. We generated approximate FSA centroids based on postal code patterns (ML), where the digit indicates north-south position and letter indicates east-west position:

$$\text{lat} = 43.65 + (\# - 5) \times 0.05 \tag{1}$$

$$\text{lon} = -79.40 + (\text{ord}(L) - \text{ord}(A)) \times 0.02 \tag{2}$$

While approximate, these coordinates provide sufficient resolution for macro-level network analysis.

4.2 Network Construction

4.2.1 Node Definition. Each of the 99 FSA areas constitutes a node $v_i \in V$ with attributes: FSA code, approximate centroid coordinates, permit count, total construction value, and temporal features.

4.2.2 Edge Construction. Nodes v_i and v_j are connected if their Euclidean distance satisfies:

$$d(v_i, v_j) = \sqrt{(\text{lat}_i - \text{lat}_j)^2 + (\text{lon}_i - \text{lon}_j)^2} \times 111 < 5 \text{ km} \tag{3}$$

The factor 111 converts degrees to kilometers. This approach yielded 165 edges with network density of 0.034, average degree of 3.33, and 12 connected components. The high clustering coefficient of 0.653 suggests meaningful spatial structure.

4.2.3 Network Metrics. Standard network metrics were computed using NetworkX: degree centrality, betweenness centrality, closeness centrality, and clustering coefficient. These metrics will serve as features for predictive models.

4.3 Feature Engineering (Planned - Week 4)

The modeling team prepared the feature engineering framework with planned features including: (1) **Accessibility**: distance to downtown, transit proximity, highway access; (2) **Amenities** from OpenStreetMap: school, park, commercial, and healthcare density; (3) **Development**: permit counts and construction value trends over 1, 2, and 5-year windows; (4) **Spatial lag**: weighted average of neighboring nodes' features where $x_i^{\text{lag}} = \frac{\sum_{j \in N(i)} w_{ij} x_j}{\sum_{j \in N(i)} w_{ij}}$; and (5) **Temporal**: historical growth rates and trend direction.

4.4 Modeling Approaches (Planned - Weeks 4-6)

4.4.1 Baseline Models. The modeling framework implements: (1) **Naive Baseline**: persistence model predicting $\Delta y_i(t+1) = \Delta y_i(t)$; (2) **LASSO Regression**: L1-regularized linear model $\min_\beta |y - X\beta|_2^2 + \lambda|\beta|_1$ for feature selection; and (3) **XGBoost**: gradient boosted decision trees for capturing non-linear relationships.

4.4.2 Spatial Models. **Spatial Autoregressive (SAR) Model**: $y = \rho W y + X\beta + \epsilon$, where ρ is the spatial autoregressive parameter and W is the spatial weight matrix. **GWR** and **GCN** are planned if time permits.

4.4.3 Evaluation Strategy. Temporal split: train on 2018-2021, validate on 2022, test on 2023. Spatial validation: train on Toronto core, test on outer GTA. Metrics: RMSE, MAE, R^2, Precision@K. Statistical tests: paired t-tests, significance testing for SAR ρ.

5 Preliminary Results

5.1 Data Collection Summary

Table 1 summarizes the data collected for this project.

Table 1: Data Collection Summary

Source	Records	Date Range	Coverage	Status
Building Permits	358,713	1981-2025	GTA-wide	Complete
FSA Areas	99	—	Toronto+	Complete

5.2 Network Properties

The constructed spatial network exhibits the following properties (Table 2): The network density of 0.034 indicates sparse connectivity, which is expected given the 5 km distance threshold and discrete FSA boundaries. The clustering coefficient of 0.653 is relatively high, suggesting that neighboring FSAs tend to form localized clusters—a desirable property for capturing neighborhood-scale spatial dependencies.

Table 2: Network Metrics

Metric	Value
Number of nodes (FSA areas)	99
Number of edges (connections)	165
Network density	0.034
Average degree	3.33
Maximum degree	5
Minimum degree	0
Is connected	False
Number of components	12
Largest component size	75 nodes
Average clustering coefficient	0.653
Avg. shortest path (largest component)	3.56

5.3 Temporal Analysis

Figure 1 shows building permits from 1981-2025 with steady growth from 1981-2000, acceleration from 2010-2019 (peak in 2016-2017), COVID-19 decline in 2020-2021, and recovery in 2022-2024. This pattern aligns with known GTA market cycles, providing confidence in data quality.

5.4 Spatial Distribution

Figure 2 visualizes the spatial network with node size representing degree centrality. The structure shows dense connectivity in downtown Toronto (M5, M6 FSAs) and sparser connectivity in outer suburbs, with multiple disconnected components representing spatially isolated regions.

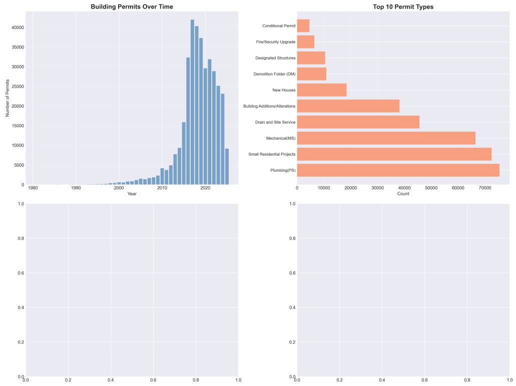

Figure 1: Building permits temporal and categorical analysis showing peak activity in 2016-2019 and COVID-19 impact in 2020-2021.

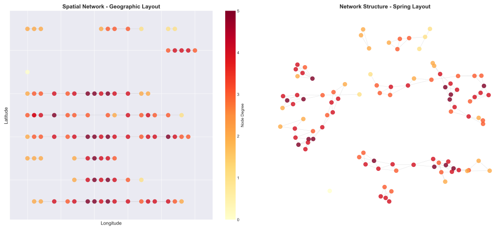

Figure 2: Spatial network visualization with 99 FSA nodes and 165 edges showing geographic and force-directed layouts.

5.5 Degree Distribution

Figure 3 shows most nodes have 2-4 connections (maximum 5), indicating homogeneous geographic connectivity rather than scale-free structure.

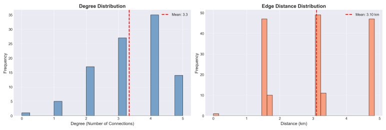

Figure 3: Network degree distribution showing majority of nodes have 2-4 connections.

5.6 Initial Insights

Preliminary results validate key assumptions: (1) publicly available building permits provide rich data for hotspot prediction, (2) 99 FSAs provide appropriate spatial resolution, (3) the network exhibits meaningful spatial structure with high clustering (0.653), and (4) clear temporal trends suggest predictable dynamics suitable for forecasting.

6 Conclusions and Future Work

6.1 Summary of Completed Work

We have successfully completed data collection (358,713 building permits), FSA aggregation (99 nodes), spatial network construction (165 edges), network metrics calculation, and visualization generation. The network exhibits realistic spatial structure with density 0.034 and clustering coefficient 0.653. Temporal analysis reveals clear development patterns aligned with known market cycles.

6.2 Challenges and Solutions

Three main challenges were addressed: (1) missing geographic coordinates solved via approximate FSA centroids from postal code patterns; (2) price data unavailability addressed by using building permits as leading indicator; and (3) NetworkX compatibility issue resolved by updating to standard pickle module.

6.3 Work Distribution and Collaboration

Weeks 1-2: Data Engineer secured sources and implemented pipeline; Network Architect researched methodologies; Modeler surveyed approaches; Analyst designed metrics. **Week 3**: Data Engineer executed collection; Network Architect implemented construction; Modeler scaffolded framework; Analyst created visualizations. Collaboration via weekly meetings, GitHub version control, and shared notebooks.

6.4 Remaining Work (Weeks 4-6)

Week 4: Integrate OpenStreetMap POI data, calculate accessibility metrics, compute spatial lag features, create temporal trend features. **Week 5**: Implement naive baseline, LASSO, and XGBoost; conduct feature importance and ablation studies. **Week 6**: Implement SAR model, test spatial spillover hypothesis, compare model performance, generate final visualizations. **Weeks 7-8**: Complete evaluation, refine visualizations, write final report and presentation.

6.5 Expected Contributions and Limitations

Upon completion, this project will demonstrate spatial network effectiveness for real estate prediction, quantify importance of accessibility, amenities, development, and spatial spillover, and provide a reproducible pipeline for FSA-level analytics using public data.

Acknowledged limitations include: FSA-level analysis with approximate coordinates, building permits as proxy for growth rather than actual prices, focus on recent years (2018-2024), and omitted macro-economic factors. These will be addressed through sensitivity analysis and clear scope definition in the final report.

6.6 Conclusion

We have successfully completed approximately 50% of project scope, establishing robust data infrastructure and network construction. Preliminary results validate our approach and demonstrate that meaningful spatial structure can be extracted from publicly available data. With clear plans for Weeks 4-6, we are well-positioned to complete the project on schedule and deliver actionable insights for GTA real estate hotspot prediction.

Author Contributions

K.W. implemented data collection infrastructure and managed data pipeline. Y.K. designed and implemented network construction algorithms. U.P. architected feature engineering and modeling framework. H.P. conducted exploratory analysis and generated visualizations. All authors contributed to experimental design, analysis interpretation, and manuscript preparation.

References

[1] Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers.

[2] Grover, A., Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. In *Proceedings of ACM KDD*.

[3] Kipf, T. N., Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of ICLR*.

[4] LeSage, J., Pace, R. K. (2009). *Introduction to Spatial Econometrics*. CRC Press.

[5] Li, Y., Yu, R., Shahabi, C., Liu, Y. (2018). Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *Proceedings of ICLR*.

[6] Wheeler, D., Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, 7(2), 161-187.

[7] Zhao, L., et al. (2020). T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21(9), 3848-3858.