Unraveling Seattle's Travel Patterns: A Data Science Exploration

Yangyang Wang, Yifei Sun

Introduction

Understanding travel patterns in urban areas like Seattle is vital for efficient transportation planning and policy-making. These patterns provide insights into how people move within the city, revealing trends in commuting, traffic congestion, and public transit usage. This project aims to analyze social demographic factors, origin-destination pairs, and transportation modes to predict travel behaviors.

This study explores Seattle's travel patterns through a data-driven approach. Insights into travel behavior and network structures were discovered by analyzing social demographic factors, origin-destination pairs, and transportation modes to predict travel behaviors and employing machine learning models and network analysis techniques. Additionally, this project develops a real-time user interface for interactive exploration. Validation with real-world data and population synthesis further enhances the robustness of the findings. This project contributes to understanding travel dynamics and informs policy-making for transportation infrastructure in urban areas.

This study also explores new ways of doing clustering on graphs using edge flow data combining both connectedness information and flow asymmetry and direction information. It compares the property of some ways of embedding nodes from edge flow data.

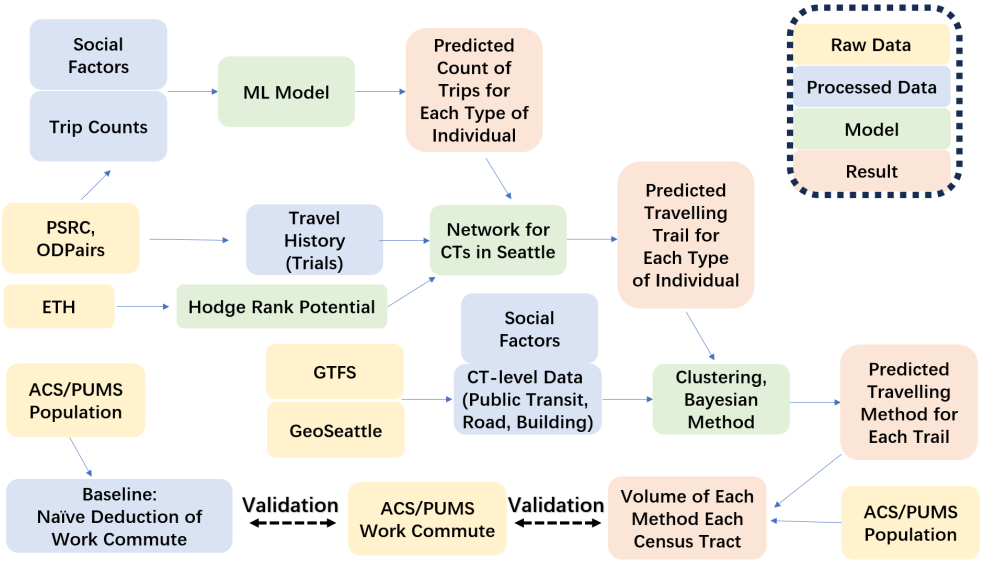


Figure 1: Panoramic Chart of This Workflow

Data

The project utilizes data from a variety of sources to analyze and understand Seattle's travel patterns. These diverse datasets provide a comprehensive view of the city's transportation dynamics and are essential for building reliable predictive models.

Puget Sound Regional Council (PSRC) Data The PSRC data includes household characteristics, population distribution, and economic factors essential for predicting trip volumes. These demographic inputs enable machine learning models to account for the impact of these variables on travel patterns and behaviors.

Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) Data LODES data, which is collected by the US Census Bureau, contains home-to-work travel patterns, aiding in mapping origin-destination pairs and uncovering commuting trends. This enables us to apply Hodge decomposition on an origin-destination travel flow graph.

UTD19 Data UTD19 data is the largest multi-city traffic dataset publically available. Collected by ETH Zurich, this dataset provides full-day traffic volume data on road networks at intervals of several minutes for 40 cities around the world. We only utilized the data of Los Angeles and Melbourne in this dataset. This dataset enables us to analyze traffic flow on a road segment network using Hodge decomposition during the day.

General Transit Feed Specification (GTFS) and Seattle's Geographical Data GTFS data provides transit schedules, routes, and stops, aiding in geographic-based clustering and method inference. Seattle's geographical data offers spatial features such as land use and topography, enabling the development of clustering models that reflect the city's unique layout.

American Community Survey (ACS) and Public Use Microdata Sample (PUMS) Data ACS and PUMS data contain social demographic information on Seattle's residents. These datasets are used for population synthesis and establishing baselines, offering a robust foundation for modeling transportation patterns and validating predictions against Work Commute data.

By leveraging these diverse data sources, the project ensures a holistic approach to understanding Seattle's travel patterns and enables the development of effective and robust transportation models.

Methodologies

Model Training and Analysis

Machine Learning Models To predict travel volume based on demographic factors, several machine learning models are employed, including Random Forest, Linear Regression, Gradient Boosting, and Deep Learning. These models use one-hot encoding to represent ten different household and individual-level features in the PSRC dataset.

Spectral clustering delineates distinct areas within King County, Seattle, based on social and geographical attributes. This process identifies factors such as demographic characteristics and spatial properties as input variables (X), with trip volume serving as the target variable (Y). Models like Random Forest and Gradient Boosting are useful for analyzing feature importance.

The project uses SHAP (SHapley Additive exPlanations) to understand the contribution of each factor towards predicting travel volume. In Figure 2, we see different factors may function their significance in different situations.

Spectral Clustering The most common motivation for spectral clustering is to minimize graph cut in the graph partitioning problem. Imagine we want to cut some edges to partition nodes into k clusters, then our goal could be to identify edges representing the weakest connections. In the first setting, we try to minimize RatioCut,

$$RatioCut\left(A_1, A_2, ..., A_k\right) = \sum_{i=1}^{k} \frac{W\left(A_i, \bar{A}_i\right)}{|A_i|}$$

where $A_1, A_2, ..., A_k$ are one partition of the nodes in the graph. For one cluster A_m, construct a normalized indicator vector $\boldsymbol{f}_{A_m\, n\times 1}$. n is the number of nodes in the graph.

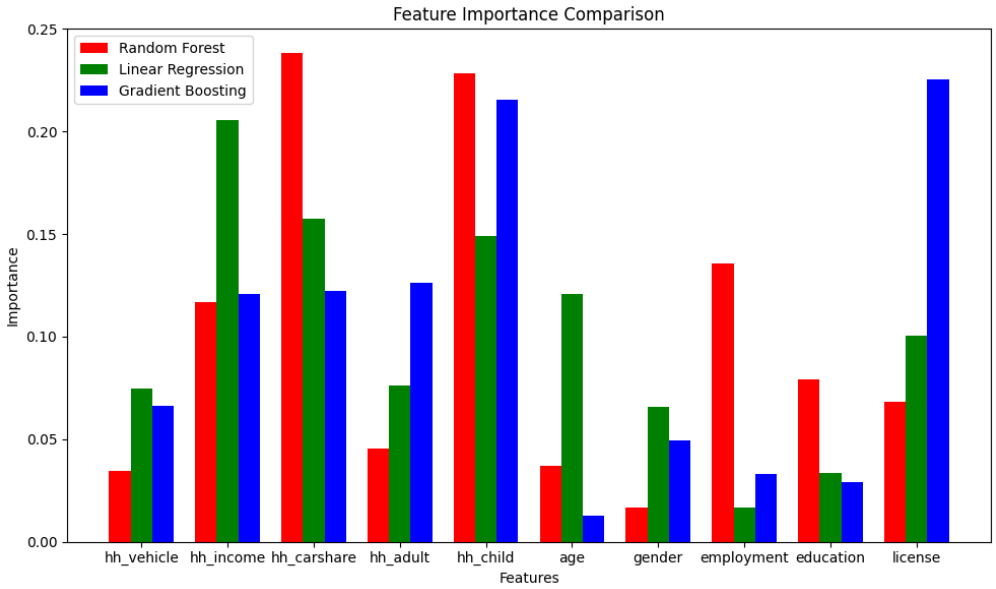


Figure 2: Feature Importance Using Various ML Models

$$\left(\boldsymbol{f}_{A_m}\right)_i = \begin{cases} 0, & \text{if } v_i \in A_m \\ \frac{1}{\sqrt{|A_m|}}, & \text{if } v_i \notin A_m \end{cases}$$

$$\boldsymbol{f}_{A_m}{}^T \boldsymbol{f}_{A_m} = |A_m| \frac{1}{\sqrt{|A_m|}} \frac{1}{\sqrt{|A_m|}} = 1$$

$$\boldsymbol{f}_{A_m}{}^T \boldsymbol{f}_{A_n} = 0$$

$$\boldsymbol{F} = \boldsymbol{F}_{n\times k} = \begin{bmatrix} \boldsymbol{f}_{A_1} & \boldsymbol{f}_{A_2} & \dots & \boldsymbol{f}_{A_k} \end{bmatrix}$$

$$\boldsymbol{F}_{n\times k}{}^T \boldsymbol{F}_{n\times k} = \boldsymbol{I}_{k\times k} = \boldsymbol{I}$$

From the property of graph Laplacian, we have

$$\boldsymbol{f}_{A_m}{}^T \boldsymbol{L} \boldsymbol{f}_{A_m} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(\left(\boldsymbol{f}_{A_m}\right)_i - \left(\boldsymbol{f}_{A_m}\right)_j \right)^2 = \frac{W\left(A_m, \bar{A_m}\right)}{|A_m|}$$

The RatioCut goal becomes

$$\operatorname{tr}(\boldsymbol{F}^T \boldsymbol{L} \boldsymbol{F}) = \operatorname{tr}(\boldsymbol{F}_{n\times k}^T \boldsymbol{L}_{n\times n} \boldsymbol{F}_{n\times k}) = \sum_{m=1}^{k} \boldsymbol{f}_{A_m}{}^T \boldsymbol{L} \boldsymbol{f}_{A_m} = \sum_{m=1}^{k} \frac{W\left(A_m, \bar{A_m}\right)}{|A_m|} = RatioCut\left(A_1, A_2, ..., A_k\right)$$

The RatioCut optimization problem can be formulated as

$$\underset{\boldsymbol{F}\in\mathbb{R}^{n\times k}}{\operatorname{argmin}} \ \operatorname{tr}\left(\boldsymbol{F}^T \boldsymbol{L} \boldsymbol{F}\right) = \operatorname{tr}\left(\boldsymbol{F}_{n\times k}{}^T \boldsymbol{L}_{n\times n} \boldsymbol{F}_{n\times k}\right) \quad s.t. \boldsymbol{F} = \boldsymbol{F}_{n\times k} = \begin{bmatrix} \boldsymbol{f}_{A_1} & \boldsymbol{f}_{A_2} & \dots & \boldsymbol{f}_{A_k} \end{bmatrix} \quad \left(\boldsymbol{f}_{A_m}\right)_i = \begin{cases} 0, & \text{if } v_i \in A_m \\ \frac{1}{\sqrt{|A_m|}}, & \text{if } v_i \notin A_m \end{cases}$$

Here the matrix $\boldsymbol{F}$ encodes the partition information of the nodes, which is similar to a one-hot encoding matrix. This matrix should be sparse. Because this problem is by nature a combinatorics problem, it is NP

hard and is difficult to get the exact optimum. We relax the constraint a bit and reformulate the problem as the following. The result of the reformulated problem has solution approximate to the original optimum. This result could be quite different from the real optimum in graph cut problem.

$$\underset{\boldsymbol{F}\in\mathbb{R}^{n\times k}}{\operatorname{argmin}}\ \operatorname{tr}\Big(\boldsymbol{F}^T\boldsymbol{L}\boldsymbol{F}\Big)=\operatorname{tr}\Big({\boldsymbol{F}_{n\times k}}^T\boldsymbol{L}_{n\times n}\boldsymbol{F}_{n\times k}\Big)\qquad s.t.{\boldsymbol{F}_{n\times k}}^T\boldsymbol{F}_{n\times k}=\boldsymbol{I}_{k\times k}=\boldsymbol{I}$$

Here the problem becomes somewhat similar to a Rayleigh quoteint problem, we just need to find a series of k orthonormal bases that minimize $\boldsymbol{F}^T\boldsymbol{L}\boldsymbol{F}$. We could just use the k eigenvectors corresponding to the k smallest non-zero eigenvalues of $\boldsymbol{L}$. The graph Laplacian $\boldsymbol{L}$ has at least one eigenvalue 0 and its associated eigenvector is all 1 vector, which is useless in partitioning the graph. Note that the matrix $\boldsymbol{F}$ we get from eigenvectors of $\boldsymbol{L}$ are neither similar to one-hot encoding nor sparse anymore.

In the second setting, we try to minimize NormalizedCut,

$$NormalizedCut\left(A_1,A_2,...,A_k\right)=\sum_{i=1}^{k}\frac{W\left(A_i,\bar{A}_i\right)}{vol(A_i)}$$

The optimization problem becomes

$$\underset{\boldsymbol{F}\in\mathbb{R}^{n\times k}}{\operatorname{argmin}}\ \operatorname{tr}\Big(\boldsymbol{F}^T\boldsymbol{L}\boldsymbol{F}\Big)=\operatorname{tr}\Big({\boldsymbol{F}_{n\times k}}^T\boldsymbol{L}_{n\times n}\boldsymbol{F}_{n\times k}\Big)\ \ s.t.\boldsymbol{F}=\boldsymbol{F}_{n\times k}=\begin{bmatrix}\boldsymbol{f}_{A_1} & \boldsymbol{f}_{A_2} & ... & \boldsymbol{f}_{A_k}\end{bmatrix}\ \ \left(\boldsymbol{f}_{A_m}\right)_i=\begin{cases}0, & \text{if } v_i\in A_m\\ \frac{1}{\sqrt{vol(A_i)}}, & \text{if } v_i\notin A_m\end{cases}$$

By relaxing the constraints, we have

$$\underset{\boldsymbol{F}\in\mathbb{R}^{n\times k}}{\operatorname{argmin}}\ \operatorname{tr}\Big(\boldsymbol{F}^T\boldsymbol{L}\boldsymbol{F}={\boldsymbol{F}_{n\times k}}^T\boldsymbol{L}_{n\times n}\boldsymbol{F}_{n\times k}\Big)\qquad s.t.{\boldsymbol{F}_{n\times k}}^T\boldsymbol{D}_{n\times n}\boldsymbol{F}_{n\times k}=\boldsymbol{I}_{k\times k}=\boldsymbol{I}$$

This is similar to a generalized Rayleigh quotient problem, we let $\tilde{\boldsymbol{F}}=\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{F}$

The optimization problem becomes

$$\underset{\tilde{\boldsymbol{F}}\in\mathbb{R}^{n\times k}}{\operatorname{argmin}}\ \operatorname{tr}\Big(\tilde{\boldsymbol{F}}^T\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}\tilde{\boldsymbol{F}}\Big)\qquad s.t.\tilde{\boldsymbol{F}}^T\tilde{\boldsymbol{F}}=\boldsymbol{I}$$

Here we just take the k eigenvectors associated to the smallest non-zero eigenvalues of the normalized graph Laplacian $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}$ as columns of $\tilde{\boldsymbol{F}}$ and we have $\boldsymbol{F}=\boldsymbol{D}^{-\frac{1}{2}}\tilde{\boldsymbol{F}}$.

For spectral clustering, there is a further step, which is to use $\boldsymbol{F}$ as embedding in Euclidean space for each node and run K-means clustering. This embedding involving the eigenvectors of graph Laplacian is called Laplacian eigenmap. The similarity matrix or adjacency matrix of the graph should be constructed wisely to reflect the similarities between nodes.

Another explanation of the spectral clustering method is in topology. The dimension of the kernel of the graph Laplacian, which is the multiplicity of eigenvalue 0, is the number of 0-holes in the graph (the number of connected components). For a well-connected graph, its non-zero eigenvalues are far from 0, while graphs with several obvious clusters will have some eigenvalues very close to 0. Eigenvectors associated with eigenvalues close to 0 indicate the almost components in the graph, and are a good way of identifying clusters in a graph.

For LODES origin-destination data, we could construct an adjacency matrix $\boldsymbol{A}$ whose entries are travel volume between a directed origin-destination pair. Because spectral clustering requires the adjacency matrix to be symmetric, we use the mean of volumes on the two opposite origin-destination pairs as the connectedness strength between the two nodes, i.e.

$$\tilde{\boldsymbol{A}}=\frac{\boldsymbol{A}+\boldsymbol{A}^T}{2}$$

We then used the Gaussian kernel to transform the average volumes on edges to pairwise similarities. Larger volume means larger similarity.

This spectral clustering of nodes on the graph does not take the direction and asymmetry of the flows into account, and only uses the average flow volume as the strength of connection between two nodes.

Hodge Decomposition on Graphs From a pure linear algebra perspective, we could have an intuitive derivation of Hodge decomposition on graphs. Figure 3 illustrates such derivation. This derivation is based on the basic facts of the four fundamental subspaces of linear transformation. The only assumption we use here is the composition of two linear transformations $\boldsymbol{A}_{m\times n}$, $\boldsymbol{B}_{n\times p}$ always gives a zero vector, i.e. $\boldsymbol{A}_{m\times n}\boldsymbol{B}_{n\times p} = \boldsymbol{0}_{m\times p}$. In topology, this composition that gives zero as its result implies that the boundary of a boundary is always empty. This assumption solely gives the result that the intermediate vector spaces (of dimension n) can be decomposed into three mutually orthogonal subspaces, i.e. $\mathrm{im}\left(\boldsymbol{B}_{n\times p}\right)$, $\ker\left(\boldsymbol{A}_{m\times n}^T\boldsymbol{A}_{m\times n} + \boldsymbol{B}_{n\times p}\boldsymbol{B}_{n\times p}^T\right)$, $\mathrm{im}\left(\boldsymbol{A}_{m\times n}^T\right)$, and when matrices $\boldsymbol{A}_{m\times n}$, $\boldsymbol{B}_{n\times p}$ are assigned to be certain meaningful matrices in the algebraic context, this decomposition becomes Hodge decomposition on graphs.

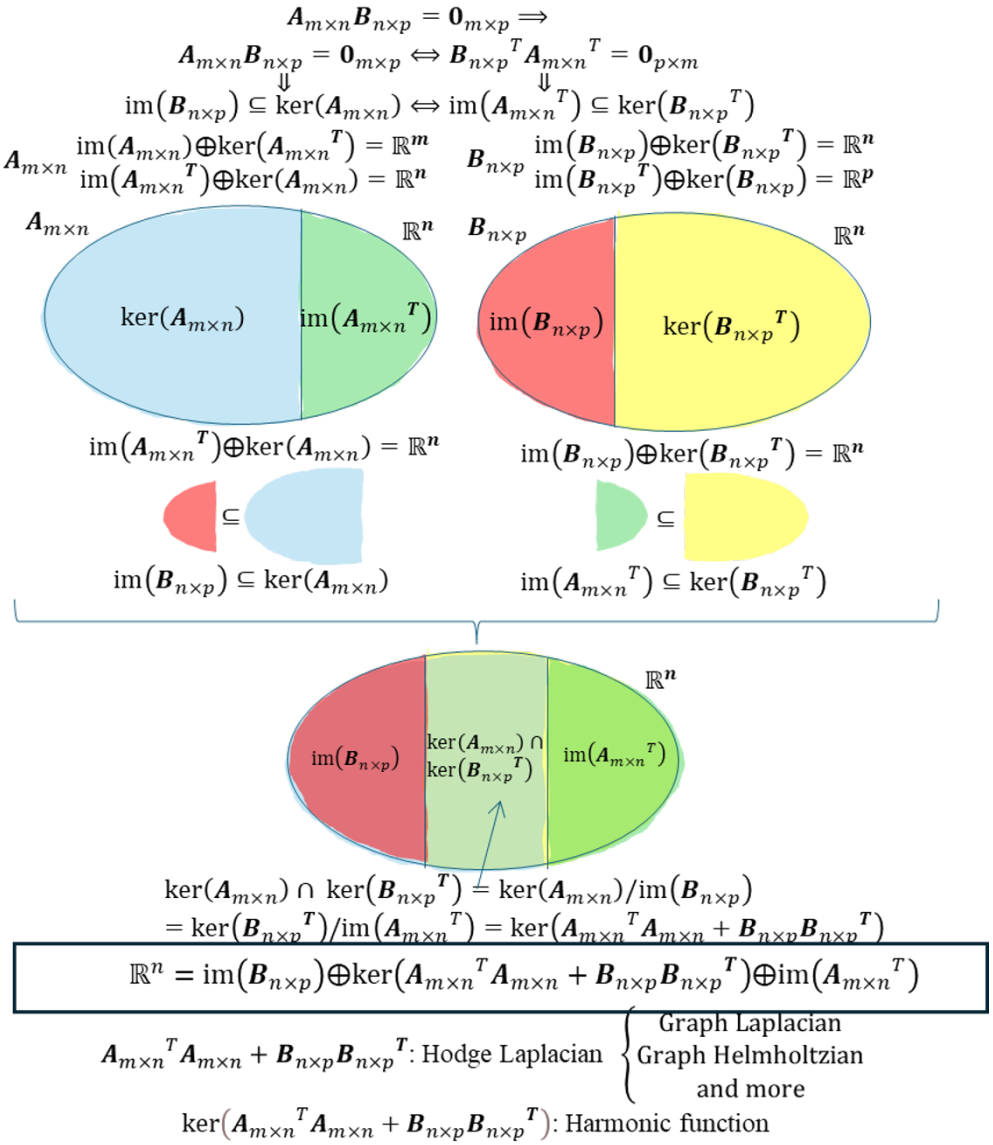


Figure 3: Illustration of the Derivation of Hodge Decomposition on Graphs (Three Mutually Orthogonal Subspaces Whose Direct Sum is the Intermediate Vector Space)

Next we need several more definitions of the operators that are meaningful in the context of discrete vector field (calculus on graphs), i.e. grad, curl, div, and curl*.

$$\boldsymbol{g} \in \mathbb{R}^{|E|}, \boldsymbol{f} \in \mathbb{R}^{|V|}, \text{grad} \in \mathbb{R}^{|E|\times|V|}, \quad \text{grad}\boldsymbol{f}_{|V|\times 1} = \boldsymbol{g}_{|E|\times 1}, \quad \boldsymbol{g}_{e_{i,j}} = (\text{grad}\boldsymbol{f})_{e_{i,j}} = f_{v_j} - f_{v_i}$$

$$\boldsymbol{g} \in \mathbb{R}^{|E|}, \boldsymbol{f} \in \mathbb{R}^{|V|}, \text{div} \in \mathbb{R}^{|V|\times|E|}, \quad \text{div}\boldsymbol{g}_{|E|\times 1} = \boldsymbol{f}_{|V|\times 1}, \quad \boldsymbol{f}_{v_i} = (\text{div}\boldsymbol{g})_{v_i} = \sum_j g_{e_{i,j}}$$

$$\boldsymbol{h} \in \mathbb{R}^{|T|}, \boldsymbol{g} \in \mathbb{R}^{|E|}, \text{curl} \in \mathbb{R}^{|T|\times|E|}, \quad \text{curl}\boldsymbol{g}_{|E|\times 1} = \boldsymbol{h}_{|T|\times 1}, \quad \boldsymbol{h}_{t_{i,j,k}} = (\text{curl}\boldsymbol{g})_{t_{i,j,k}} = g_{e_{i,j}} + g_{e_{j,k}} + g_{e_{k,i}} = g_{e_{i,j}} - g_{e_{i,k}} + g_{e_{j,k}}$$

$$-\text{div} = \text{grad}^*$$

$$\text{grad} \circ \text{curl} = 0, \quad \text{curl}^* \circ -\text{div} = 0$$

These four operators in this discrete context are in matrix form. They are some low-dimension special cases of boundary operators and coboundary operators in algebraic topology. Hodge decomposition is the decomposition of cochains defined on any dimensional simplicial complex (functions defined on k-cliques). The boundary relationship between k-simplicial complex and k+1-simplicial complex is described by the operator operator. grad and $-$div are the adjoints (Hermitian conjugates) of each other, and are the lowest order boundary operators. They are merely the incidence matrix and its transpose in graph theory. The composition of grad, and curl and composition of $-$div, and curl* always give the result of zero, which can be easily verified. This unveils the more fundamental result that the boundary of a boundary is zero.

In general, cochains are functions defined on simplices $X \to \mathbb{R}$, for k-simplicial complex, they can be represented as a vector, i.e. $\mathbb{R}^{|V|}$, $\mathbb{R}^{|E|}$, $\mathbb{R}^{|T|}$, etc. The sign of the value of the function is related to the parity of the permutation.

$$f\left(\left[i_{p(0)}, ..., i_{p(k)}\right]\right) = \text{sign}(p) f([i_0, ..., i_k])$$

$i_{p(0)}, ..., i_{p(k)}$ is a permutation of $i_0, ..., i_k$. If it is an odd permutation, $\text{sign}(p) = -1$; if it is a even permutation, $\text{sign}(p) = 1$

Coboundary operator δ_k is the linear map from lower order cochains to higher order cochains and is defined as

$$(\delta_k f)(i_0, i_1, ..., i_{k+1}) := \sum_{j=0}^{k+1} (-1)^j f(i_0, ..., i_{j-1}, i_{j+1}, ..., i_{k+1})$$

We can see the above defined operators are merely special cases of the definitions here.

From the above derivation we have

$$\boldsymbol{A}_{m\times n}\boldsymbol{B}_{n\times p} = \boldsymbol{0}_{m\times p} \longrightarrow \mathbb{R}^n = \text{im}\left(\boldsymbol{B}_{n\times p}\right) \oplus \ker\left(\boldsymbol{A}^T_{m\times n}\boldsymbol{A}_{m\times n} + \boldsymbol{B}_{n\times p}\boldsymbol{B}^T_{n\times p}\right) \oplus \text{im}\left(\boldsymbol{A}^T_{m\times n}\right)$$

By limiting the context to node space, edge space and triangle space, we could replace $\boldsymbol{A}_{m\times n}$, $\boldsymbol{B}_{n\times p}$, $\boldsymbol{A}^T_{m\times n}$, and $\boldsymbol{B}^T_{n\times p}$ with curl, grad, curl*, and $-$div and then get

$$\begin{aligned}\text{curl}\ \ \text{grad} = \boldsymbol{0}_{m\times p} \longrightarrow \mathbb{R}^{|E|} &= \text{im}\left(\text{grad}\right) \oplus \ker\left(\text{curl}^* \circ \text{curl} + \text{grad} \circ (-\text{div})\right) \oplus \text{im}\left(\text{curl}^*\right) \\ &= \text{im}\left(\text{grad}\right) \oplus \ker\left(-\text{grad} \circ \text{div} + \text{curl}^* \circ \text{curl}\right) \oplus \text{im}\left(\text{curl}^*\right)\end{aligned}$$

$$\text{graph}\ \ \text{Laplacian} : -\text{div} \circ \text{grad}$$

$$\text{graph}\ \ \text{Helmholtzian} : -\text{grad} \circ \text{div} + \text{curl}^* \circ \text{curl}$$

$\ker\left(-\text{grad} \circ \text{div} + \text{curl}^* \circ \text{curl}\right)$ is called the 2nd homology group, which are cycles (which have empty boundary) but not the boundary of anything else. In other words, they are cycles longer than 3. (Cycles of

length 3 would be the boundary of an triangle.) $\dim(\ker(-\text{grad} \circ \text{div} + \text{curl}^* \circ \text{curl}))$ is the number of order 2 holes in the graph, also called Betti number.

Figure 4 is an illustration of the Hodge decomposition we used in the project. It is a discrete analogy of the Hodge decomposition of a continuous vector field. Here we define three spaces on the graph, which are node space, edge space, and triangle space. In algebraic topology, nodes, edges, and triangles (1-clique, 2-clique, and 3-clique in graph theory) correspond to 0-simplicial complex, 1-simplicial complex, and 2-simplicial complex. Every node, edge, and triangle (0-simplex, 1-simplex, 2-simplex) can have a value defined on it, and thus create three vector spaces associated with nodes, edges, and triangles respectively. The relation between cliques in graphs and simplices in topology is shown in Figure 5. Boundary operators and their adjoints define the relationship between those three spaces.

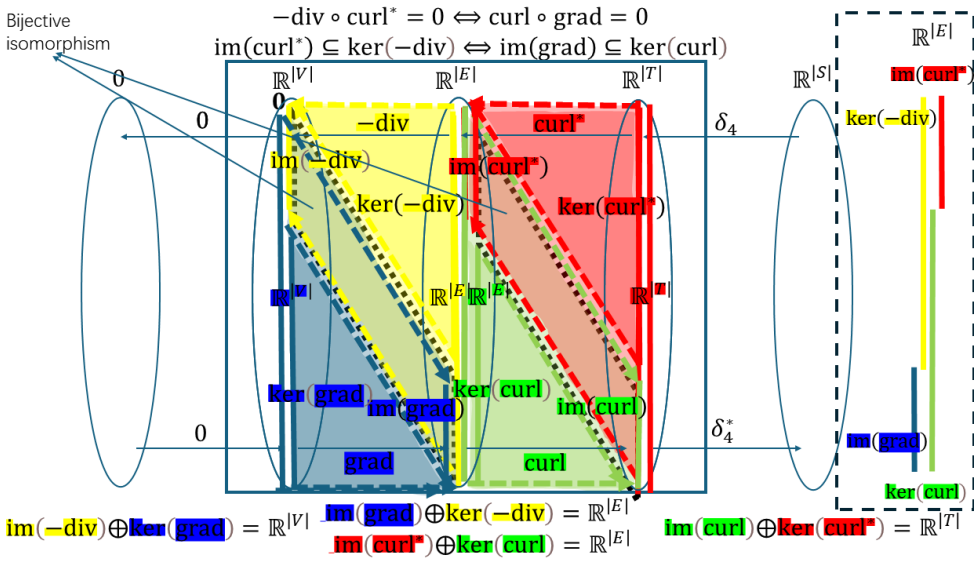


Figure 4: Illustration of Hodge Decomposition

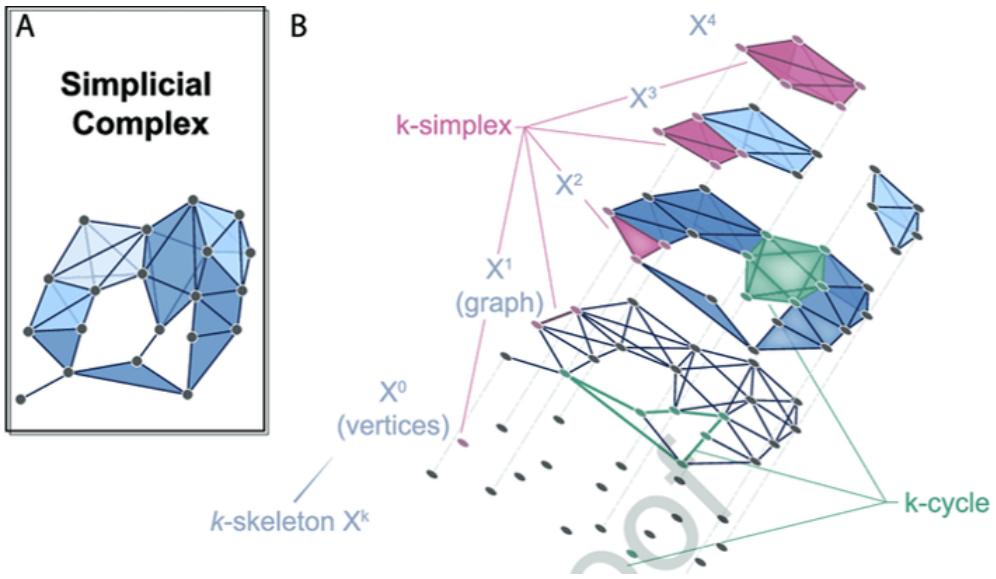


Figure 5: Illustration of Simplex and Simplicial Complex in Graphs

These three components have respective meanings, i.e.

$$\text{Edge Flow} = \overbrace{\text{conservative} \oplus \underbrace{\text{solenoidal and irrotational}}_{\text{solenoidal}} \oplus \text{vorticity}}^{\text{irrotational}}$$

$$\text{Edge Flow} = \overbrace{\text{curl} - \text{free but not divergence} - \text{free} \oplus \underbrace{\text{divergence} - \text{free and curl} - \text{free} \oplus \text{divergence} - \text{free but not curl} - \text{free}}_{\text{divergence-free}}}^{\text{curl-free}}$$

In HodgeRank, only the $\text{im}\,(\text{grad})$ (conservative) component is used to calculate the ranking values of nodes. For $\text{im}\,(\text{grad})$ component of the edge flow, the flow value on every edge (predefined with an orientation) is the difference between the values of the two end nodes. The values defined on nodes that create this edge flow component is called the potential of this graph. In this sense, $\text{im}\,(\text{grad})$ is perfectly consistent for every pair of neighboring nodes. $\text{im}\,(\text{curl}^*)$ is used as the measurement for local inconsistency (inconsistency between three mutually connected nodes). $\ker\,(-\text{grad} \circ \text{div} + \text{curl}^* \circ \text{curl})$ is used as measurement for global inconsistency (inconsistency on cycles longer than 3). Larger $\text{im}\,(\text{grad})$ component in the edge flow indicates the high rankability of the graph, and vice versa.

To get the potentials on the nodes of a graph given the edge flows, we would just project the edge flows onto the node space in order to get the potential vector that could create the closest purely gradient edge flows to the raw edge flows (because of the orthogonality condition). This is similar to solving linear regression using least squares, in which $\boldsymbol{y}$, $\boldsymbol{X}$, and β in $\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$ are replaced by $\boldsymbol{f}$ (edge flows defined on edges), grad (gradient operator), and $\boldsymbol{r}$ (potentials defined on nodes) as illustrated in Figure 6. The difference is that in linear regression we assume independent columns in the matrix $\boldsymbol{X}$, and we try to avoid multicollinearity. However, this incidence matrix grad is always neither full column rank nor full row rank, so there are infinite approximate solutions by projection, and these solutions are addition invariant, which makes sense because potential is only meaningful when there is a reference point. Another feature of the matrix grad is that it is very sparse, therefore we used scipy.sparse_matrix to accelerate computation. Also, this matrix could have a lot more edges than nodes if we assume the graph to be similar to an Erdos-Renyi random graph, because the number of edges is quadratic to the number of nodes in this setting, meaning that this matrix is super tall and slim.

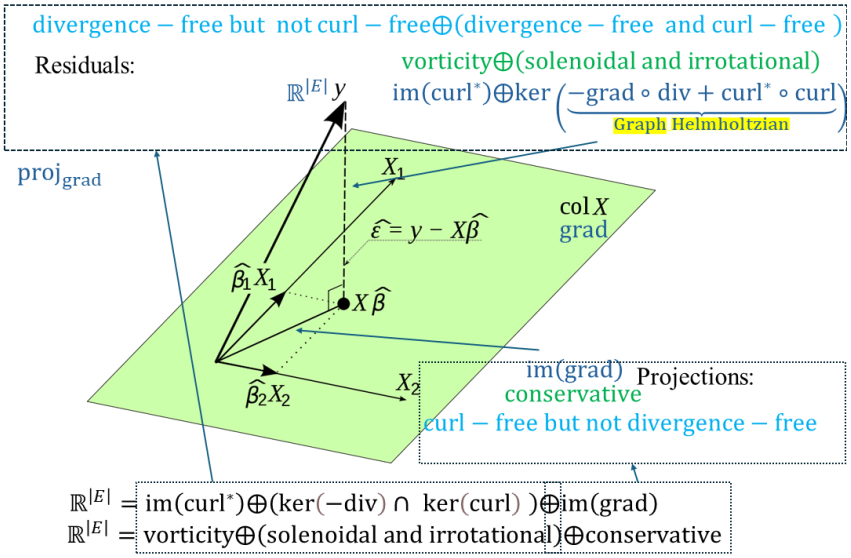


Figure 6: Illustration of HodgeRank Using Projection

To solve this we would normally employ the Moore-Penrose pseudo-inverse of matrix grad (here denoted as ∂_1^T, meaning the adjoint of boundary operator of order 1). Note in HodgeRank, there could be a weight assigned to each edge, and the linear regression is solved using weighted least squares.

$$\min_{\boldsymbol{r}\in\mathbb{R}^{|V|}} \|\boldsymbol{f}-\partial_1^T\boldsymbol{r}\|_{\boldsymbol{W}}^2$$

The solutions are

$$\boldsymbol{r}=\left(\partial_1\boldsymbol{W}^{\frac{1}{2}}\right)^{T^\dagger}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{f}+\left(\boldsymbol{I}-\left(\partial_1\boldsymbol{W}^{\frac{1}{2}}\right)^{T^\dagger}\left(\partial_1\boldsymbol{W}^{\frac{1}{2}}\right)^{T}\right)\boldsymbol{w}$$

For UTD19 data, we have the traffic volumes in both directions of a road segment. The first step we need to do is to construct the adjacency matrix $\boldsymbol{A}$, whose entries are the traffic volume from the start node to the end node of a road segment. We will define the direction of each road segment and get the difference of volumes in both directions, i.e.

$$\tilde{\boldsymbol{A}}=\boldsymbol{A}-\boldsymbol{A}^T$$

Note that here an edge of net flow 0 does not have the same effect on the network as that edge does not exist. The nonexistence of an edge between two nodes indicates there is no possible information on the difference between the potentials of the two nodes, while an edge with the same volumes on both directions indicates their potentials are somewhat the same. The structure of the network and the flow values on the network are two different inputs of the model.

Our motivation for using HodgeRank on this traffic flow network is to model it like an electric circuit. People travel in the network from nodes to nodes like electrons travel in a circuit from nodes to nodes. There is a potential distributed across the nodes in the network that drives people to commute from suburbs to downtown during morning rush hour and reversely during evening rush hour. Here the inherent assumption is that the resistance on every road is the same value 1, so the current and voltage on a road segment are the same.

The effect of HodgeRank method is that when given all edge flows, we could get the potentials on nodes and how confident the potentials are. It achieves an embedding of nodes using edge data.

Bayesian Methods for Transportation Mode Prediction Bayesian methods are utilized to predict transportation mode choices based on conditional probabilities. Specifically, Naive Bayes models and Partial Least Squares Regression (PLSR) are employed. These methods are preferred because they handle the high number of factors involved and accommodate the addition of more variables over time.

Results

Demographic Analysis

The project presents findings on how social demographic factors influence travel behavior in Seattle. By visualizing model predictions, the accuracy of travel volume predictions for different demographic groups can be assessed. These insights help understand the impact of population characteristics on transportation choices and behaviors.

Spectral Clustering

We use the the average of bidirectional travel volumes between two nodes in the PSRC as connection strength on that edge and cluster the graph to 11 clusters. Because the clusters in the result become very disconnected, we add the distance data into the similarity matrix to assign closer nodes with higher similarity. The result is shown in Figure 5. From the results in Appendix B, we can see that increasing weight of the distance factor in the similarity matrix will reduce the disconnectedness in the clusters. The inherent paradox here is that geographical proximity and traveling volume could be mutually exclusive. People may tend to travel to places that are far from their origins. There could be further analysis on traveling volume and travel distance distribution. The criteria for similarity between two nodes in a flow graph could be further explored, for example, average traveling volume, asymmetry in the traveling volume in terms of direction, and geographical proximity. That is part of the reason why we employed HodgeRank in this study.

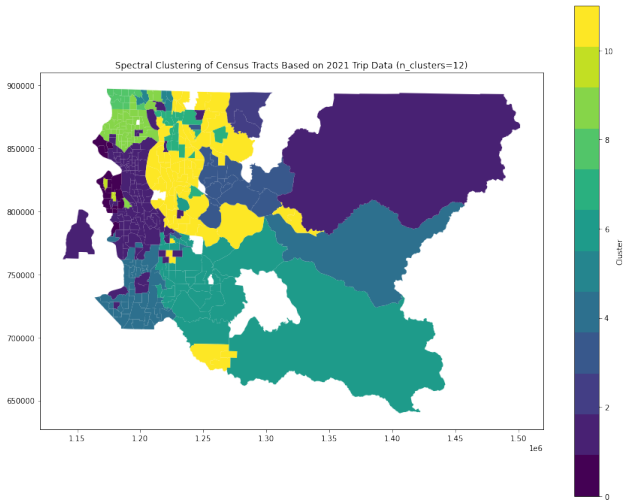


Figure 7: Spectral Clustering of Census Tracts in King County, Washington Based on PSRC Data

HodgeRank Negative Potential and Divergence

The negative potential in King County, Washington from HodgeRank using LODES data is shown in Figure 8. Highlighted areas of high negative potential is where people tend to travel for work, and darker areas indicate residential area.

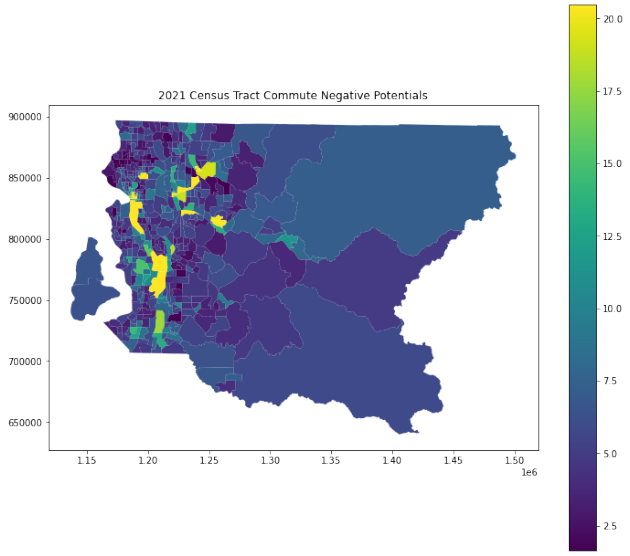


Figure 8: Negative Potential Values From HodgeRank of King County, Washington Using Work-Home Data From LODES Data

In this Hodge decomposition framework, two functions can be defined on nodes, which are potential and divergence. Potential indicates the high or low position of nodes in this flow network. Divergence is simply the net inflow minus outflow in the nodes, indicating the sheer volume into or out of nodes. They are related through the following equations

$$\boldsymbol{f}_{|E|\times 1} = \boldsymbol{M}_{|E|\times|V|}\boldsymbol{p}_{|V|\times 1}$$

$$\boldsymbol{d}_{|V|\times 1} = \boldsymbol{M}^{T}_{|E|\times|V|}\boldsymbol{f}_{|E|\times 1}$$

$$\boldsymbol{d}_{|V|\times 1} = \boldsymbol{M}^T_{|E|\times|V|}\boldsymbol{M}_{|E|\times|V|}\boldsymbol{p}_{|V|\times 1} = \boldsymbol{L}_{|V|\times|V|}\boldsymbol{p}_{|V|\times 1}$$

$$\boldsymbol{p}_{|V|\times 1} = \boldsymbol{L}^{\dagger}_{|V|\times|V|}\boldsymbol{d}_{|V|\times 1}$$

$\boldsymbol{f}_{|E|\times 1}$ is gradient flows on edges, $\boldsymbol{p}_{|V|\times 1}$ is potentials on nodes, $\boldsymbol{d}_{|V|\times 1}$ is divergences on nodes. $\boldsymbol{M}_{|E|\times|V|}$ is the directed incidence matrix. When there is at most one edge between every pair of nodes on the directed graph, $\boldsymbol{M}^T_{|E|\times|V|}\boldsymbol{M}_{|E|\times|V|}$ is equal to the graph Laplacian $\boldsymbol{L}_{|V|\times|V|}$. In the LODES dataset, we noticed potential and divergence of nodes tend to have a piecewise linear relationship with 0 as its inflection point, as shown in Figure 9.

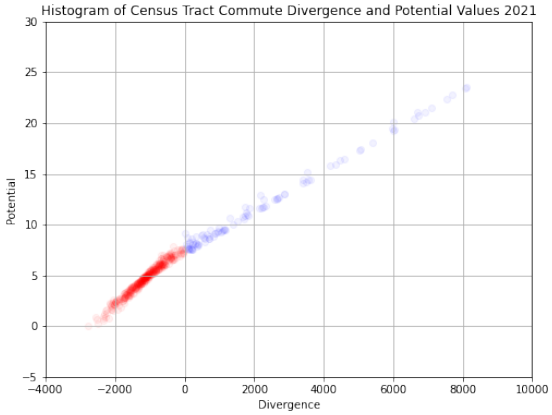


Figure 9: Negative Potential Values and Divergence of King County, Washington Using Work-Home Data From LODES Data

The reason for this piecewise linear relationship is that this network constructed from origin-destination data has some particular features. Nodes in the download area have very large degrees and receive large influx during work time, while nodes in suburbs have small degrees and witness net outflow during work time. The graph Laplacian of this network is almost diagonal with diagonal entries at the magnitude of several hundred and non-diagonal entries very sparse and being 0 or -1 (see Appendix C). The diagonal entries of the pseudo inverse of the graph Laplacian are very close to the inverses of diagonal entries of the graph Laplacian. Therefore the slopes for divergence > 0 and divergence < 0 are different. Also, this high linearity indicates that using the potential to rank nodes in the network becomes almost meaningless, as it roughly gives the same ranking as using divergence, and divergence is easier to compute as it only involves doing matrix multiplication while potential requires matrix inversion. This could be a complementary conclusion for [1] as we used the same dataset and conducted the same tasks.

Because this network only contains origin-destination information, nodes far away could be connected, resulting in high degrees for downtown nodes. This network also ignores the details of trips as people can not teleport from origins to destinations. Therefore, we used a new dataset, UTD19, to apply HodgeRank on a real road network in downtown Los Angeles. Because of the mesh-like shape of the graph, this graph has much fewer edges than the origin-destination network. The results are shown in Appendix D.

We can see that during morning and evening rush hours, the high potential areas and lower potential areas are reversed. As people move from high-potential areas to lower-potential areas, these plots indicate the commute corridor of downtown LA is from downtown to northeast and southwest. From the scatter plots, we notice that, unlike the origin-destination network, potential and divergence here are slightly positively correlated but the significance is much weaker. This is because the nodes in the network have degrees of at most 5, and the non-diagonal entries in the pseudo inverse of the graph Laplacian have values of large magnitudes. And there is not a significant difference in degrees between different nodes. From the histogram, we can see that potential and divergence are of roughly the same magnitude range, but potential distribution is more spread out while divergence distribution is more centered around 0. We can see that the values

of potential change smoothly across neighboring nodes in the graph while the values of divergence change abruptly across neighboring nodes. To further investigate this property, we develop the concepts of local variance and global variance on graphs.

The sample variance for a sequence of values is

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

When expressed in vector form,

$$\frac{1}{n-1}\boldsymbol{x}^T\left(\boldsymbol{I}-\frac{1}{n}\mathbf{1}_{n\times n}\right)\boldsymbol{x}$$

$n\left(\boldsymbol{I}-\frac{1}{n}\mathbf{1}_{n\times n}\right)=n\boldsymbol{I}-\mathbf{1}_{n\times n}$ can be interpreted as the graph Laplacian of a complete graph of n nodes.

The quadratic form of a graph Laplacian can describe the sum of squared differences of neighboring nodes' values

$$\boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x}=\frac{1}{2}\sum_{i\sim j}(x_i-x_j)^2$$

We define

$$\text{local var}(\boldsymbol{x})=\boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x}\qquad\text{global var}(\boldsymbol{x})=\boldsymbol{x}^T\left(\boldsymbol{I}-\frac{1}{n}\mathbf{1}_{n\times n}\right)\boldsymbol{x}$$

Through some calculation and approximation, we could get the expression of global variance and local variance of potential and divergence.

$$\text{global var}(\boldsymbol{p})\approx\boldsymbol{f}^T\boldsymbol{U}_{|E|\times|E|}\begin{bmatrix}\Lambda_{|V|\times|V|}^{-1} & \mathbf{0}_{|V|\times(|E|-|V|)}\\ \mathbf{0}_{(|E|-|V|)\times|V|} & \mathbf{0}_{(|E|-|V|)\times(|E|-|V|)}\end{bmatrix}\boldsymbol{U}_{|E|\times|E|}{}^T\boldsymbol{f}=\sum_{i=1}^{|V|}\frac{a_i^2}{\lambda_i}$$

$$\text{global var}(\boldsymbol{d})\approx\boldsymbol{f}^T\boldsymbol{U}_{|E|\times|E|}\begin{bmatrix}\Lambda_{|V|\times|V|} & \mathbf{0}_{|V|\times(|E|-|V|)}\\ \mathbf{0}_{(|E|-|V|)\times|V|} & \mathbf{0}_{(|E|-|V|)\times(|E|-|V|)}\end{bmatrix}\boldsymbol{U}_{|E|\times|E|}{}^T\boldsymbol{f}=\sum_{i=1}^{|V|}\lambda_i a_i^2$$

$$\text{local var}(\boldsymbol{p})=\boldsymbol{f}^T\boldsymbol{U}_{|E|\times|E|}\begin{bmatrix}\boldsymbol{I}_{|V|\times|V|} & \mathbf{0}_{|V|\times(|E|-|V|)}\\ \mathbf{0}_{(|E|-|V|)\times|V|} & \mathbf{0}_{(|E|-|V|)\times(|E|-|V|)}\end{bmatrix}\boldsymbol{U}_{|E|\times|E|}{}^T\boldsymbol{f}=\sum_{i=1}^{|V|}a_i^2$$

$$\text{local var}(\boldsymbol{d})=\boldsymbol{f}^T\boldsymbol{U}_{|E|\times|E|}\begin{bmatrix}\Lambda_{|V|\times|V|}^{2} & \mathbf{0}_{|V|\times(|E|-|V|)}\\ \mathbf{0}_{(|E|-|V|)\times|V|} & \mathbf{0}_{(|E|-|V|)\times(|E|-|V|)}\end{bmatrix}\boldsymbol{U}_{|E|\times|E|}{}^T\boldsymbol{f}=\sum_{i=1}^{|V|}\lambda_i{}^2 a_i^2$$

$\boldsymbol{U}$, Λ and $\boldsymbol{V}$ are results from SVD of incidence matrix. $\boldsymbol{a}=\boldsymbol{U}_{|E|\times|E|}{}^T\boldsymbol{f}$. These four values can all be expressed in a weighted sum of squares.

We can see the properties of these four quantities are mostly determined by the singular values of the incidence matrix (or the eigenvalues of the graph Laplacian). From Appendix D, we can see that some eigenvalues of graph Laplacian are close to 0, while the majority of it is greater than 1. One reasonable explanation is that the squares of eigenvalues in the local variance of divergence make it a lot larger than the local variance of potential. And inverses of some eigenvalues very close to 0 make the global variance of potential larger than the global variance of divergence. The experiment results are shown in Figure 10.

This tells us the potential is better than divergence in embedding nodes in the flow network. It is able to better differentiate different nodes globally and also embed neighboring nodes to similar values in the result. The assortativities of potential and divergence in Figure 11 also show that potential values of neighboring nodes tend to be similar while divergence value of neighboring nodes tend to differ a lot.

Finally, we propose a new idea to do clustering on a flow graph. There should be three factors that need to be incorporated into this clustering process: the structure of the network, the mean flow of both directions on every edge, and the net difference of flows in both directions on every edge. The intuition is

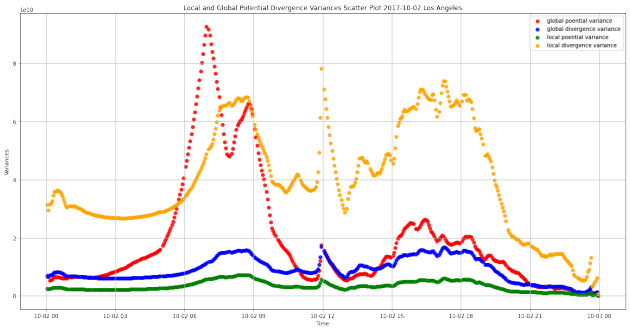


Figure 10: The local and global variances for potential and divergence in a Day in LA

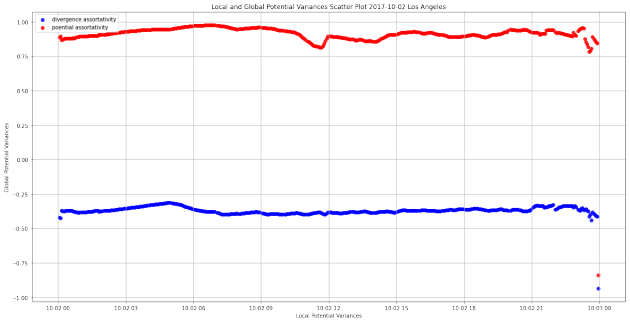


Figure 11: The assortativities for potential and divergence in a Day in LA

nodes that are connected by edges should be similar, nodes that have a large flow on edges between them should be similar, and nodes that have a small net flow difference on edges between them should be similar. This would be a modification of spectral clustering. Spectral clustering is just first doing Laplacian eigenmap and then using K-means. Here we add more information to the embedding step to incorporate volume and volume differences on edges. The results are shown in Appendix E. Clustering based on potential embedding creates layered results, which can be rectified by adding network structure information to it.

Transportation Mode Prediction

Bayesian model results show the probabilities of using different transportation modes for trips between census tracts. Comparing predicted mode choices with ground truth data provides insight into the accuracy and reliability of the models. This analysis supports better decision-making regarding transportation policies and infrastructure.

Real-time Visualization and User Interface

The project includes the development and deployment of a user-friendly web interface that allows users to input individual attributes and starting points. Backend algorithms generate trip volume, likely destinations, and probabilities of transportation methods for each trip.

The integration of the model with Flask and Gunicorn ensures real-time deployment, scalability, and responsiveness. User testing and feedback collection evaluate the usability and effectiveness of the interface.

Validation

Validation involves combining the population of individuals in each census tract to derive transportation volume and comparing it against a naive approach based on population proportion. This validation is also conducted against Work Commute data.

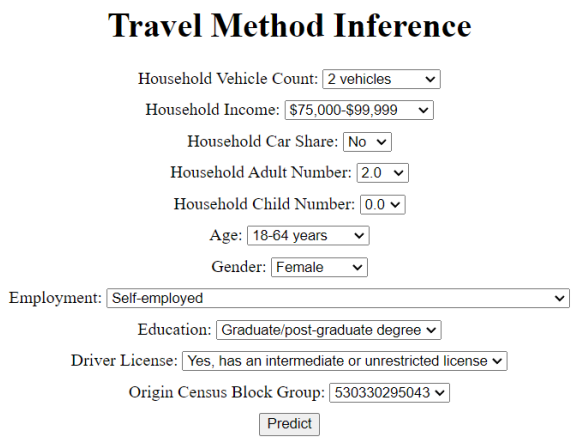
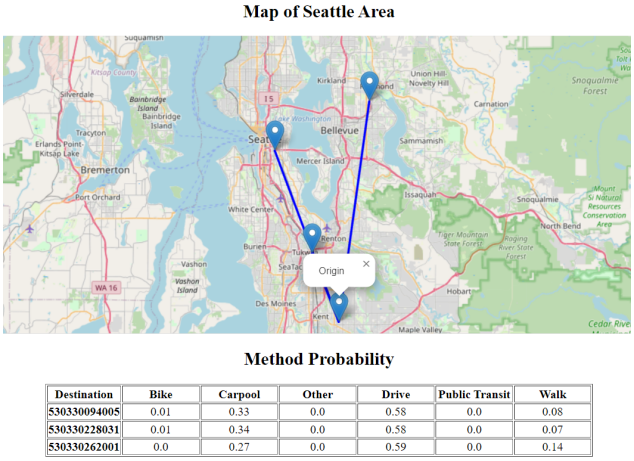


Figure 12: Input Boxes in Our Website



Destination	Bike	Carpool	Other	Drive	Public Transit	Walk
530330094005	0.01	0.33	0.0	0.58	0.0	0.08
530330228031	0.01	0.34	0.0	0.58	0.0	0.07
530330262001	0.0	0.27	0.0	0.59	0.0	0.14

Figure 13: Result Outputs Once Clicked 'Predict'

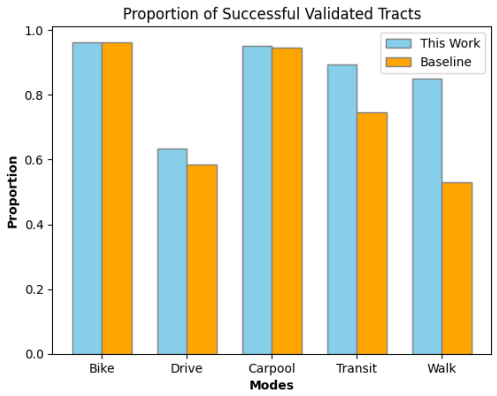


Figure 14: Proportion of Correctly Predicted Census Tracts

The project's significance lies in its accuracy improvement, granularity enhancement, and policy insights. It significantly outperforms baseline methods in detecting low accuracy in transit, drive, and walk modes which can be deduced from 14. The model provides finer granularity, offering data specific to census block groups, unlike the coarse census tract data in ACS.

Discussion

The findings have important implications for transportation planning in Seattle. They highlight the influence of social demographic factors on travel behaviors and the potential for improving transportation infrastructure and policies. The project addresses limitations such as data biases, model assumptions, and uncertainties.

Suggestions for future research include incorporating real-time data and dynamic modeling approaches. These enhancements could further improve the accuracy and applicability of the models for decision-making.

Analysis of properties of potential and divergence reveals possible applications in clustering and embedding in directed flow graphs. More studies could be done on how to define similarity and better cluster in flow graphs.

Conclusion

This project contributes to understanding travel behavior in Seattle and its potential for informing policy and infrastructure decisions. By combining data science, transportation planning, and urban geography, the study provides valuable insights for city planners and decision-makers. The project's findings offer a foundation for future research and development in transportation planning and urban mobility.

References

[1] Takaaki Aoki, Shota Fujishima, and Naoya Fujiwara. "Urban spatial structures from human flow by Hodge–Kodaira decomposition". In: *Scientific reports* 12.1 (2022), p. 11258.

[2] José Carpio-Pinedo, Manuel Benito-Moreno, and Patxi J Lamíquiz-Daudén. "Beyond land use mix, walkable trips. An approach based on parcel-level land use data and network analysis". In: *Journal of Maps* 17.1 (2021), pp. 23–30.

[3] Lei Gong, Ryo Kanamori, and Toshiyuki Yamamoto. "Data selection in machine learning for identifying trip purposes and travel modes from longitudinal GPS data collection lasting for seasons". In: *Travel Behaviour and Society* 11 (2018), pp. 131–140.

[4] Gabriel Goulet-Langlois et al. "Measuring regularity of individual travel patterns". In: *IEEE Transactions on Intelligent Transportation Systems* 19.5 (2017), pp. 1583–1592.

[5] Julian Hagenauer and Marco Helbich. "A comparative study of machine learning classifiers for modeling travel mode choice". In: *Expert Systems with Applications* 78 (2017), pp. 273–282.

[6] Zhao Ji. *Hodgerank: Generating movie ranking from IMDb movie ratings, part 1*. Accessed: April 16, 2024. Jan. 15, 2022. URL: `https://medium.com/@zj444/hodgerank-generating-movie-ranking-from-imdb-movie-ratings-part-1-2a88ec148f10`.

[7] Xiaoye Jiang et al. "Statistical ranking and combinatorial Hodge theory". In: *Mathematical Programming* 127.1 (2011), pp. 203–244.

[8] Lek-Heng Lim. "Hodge Laplacians on graphs". In: *Siam Review* 62.3 (2020), pp. 685–715.

[9] Sean Peirce and Jane E Lappin. *Acquisition of traveler information and its effects on travel choices: evidence from a Seattle-area travel diary survey*. Tech. rep. Citeseer, 2003.

[10] Dominic Stead and Stephen Marshall. "The relationships between urban form and travel patterns. An international review and evaluation". In: *European journal of transport and infrastructure research* 1.2 (2001).

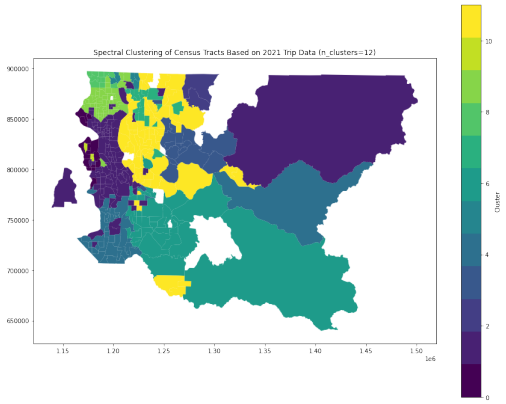
[11] Junshi Xu, Marc Saleh, and Marianne Hatzopoulou. "A machine learning approach capturing the effects of driving behaviour and driver characteristics on trip-level emissions". In: *Atmospheric environment* 224 (2020), p. 117311.

Appendix A

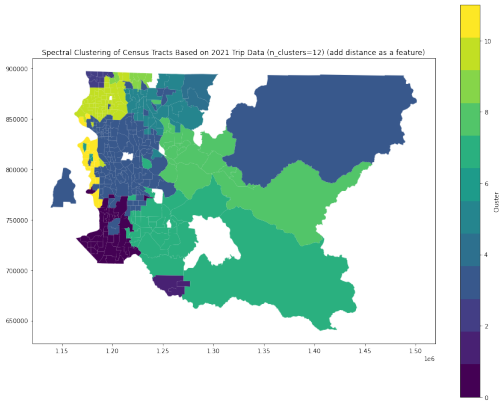
For further inquiries or to access additional resources, please use the following contact information and links:

- Email: wyy@umich.edu, yifeisun@umich.edu
- GitHub Repository: wyy-frank/SI699-Repo
- Website: seattle-travel.replit.app

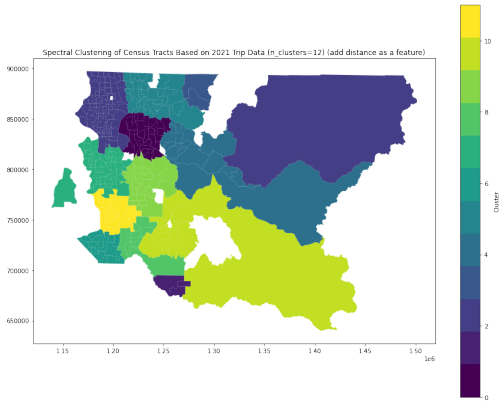
Appendix B



(a) Spectral Clustering of Census Tracts in King County, Washington Based on PSRC Data

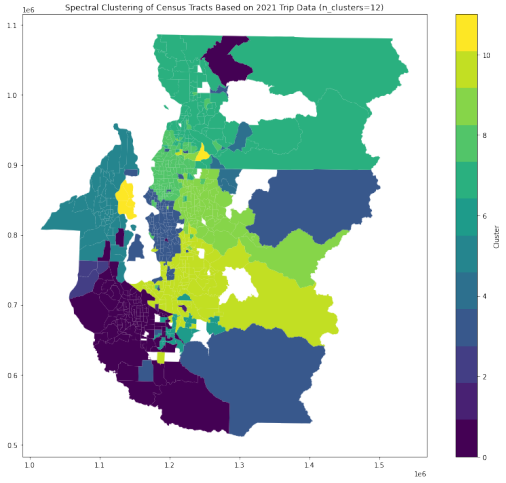


(b) Spectral Clustering of Census Tracts in King County, Washington Based on PSRC Data and Distance Data

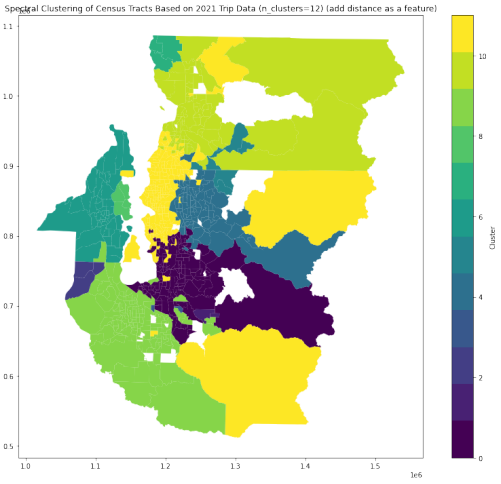


(c) Spectral Clustering of Census Tracts in King County, Washington Based on PSRC Data and Distance Data (Larger Weight for Distance)

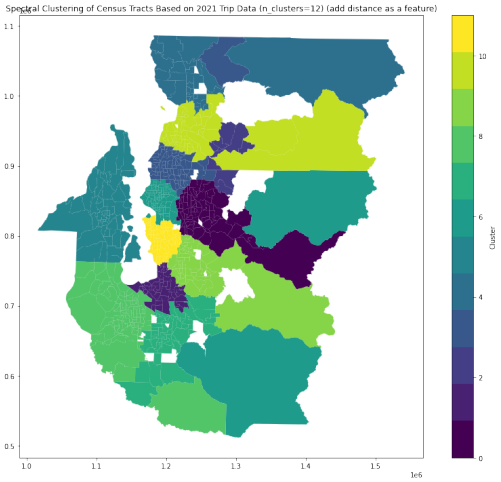
Figure 15: Spectral Clustering of Census Tracts in King County, Washington Based on PSRC Data



(a) Spectral Clustering of Census Tracts in Puget Sound Based on PSRC Data



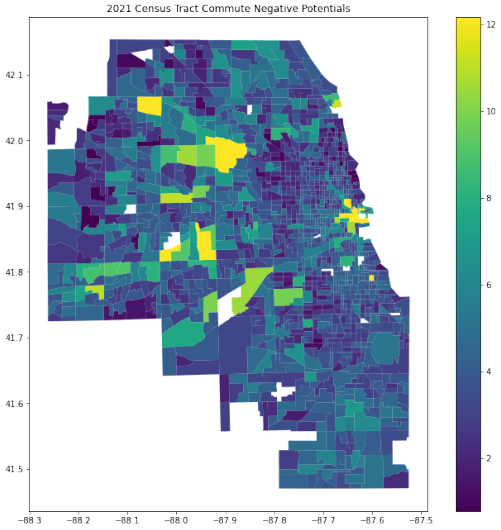
(b) Spectral Clustering of Census Tracts in Puget Sound Based on PSRC Data and Distance Data



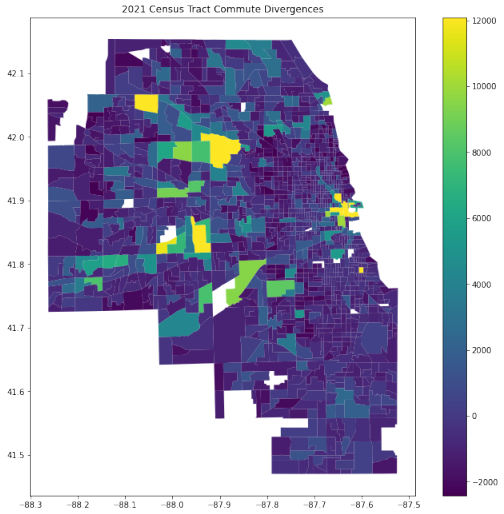
(c) Spectral Clustering of Census Tracts in Puget Sound Based on PSRC Data and Distance Data (Larger Weight for Distance)

Figure 16: Spectral Clustering of Census Tracts in Puget Sound Based on PSRC Data

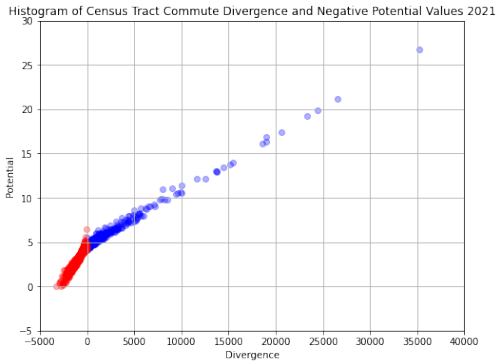
Appendix C



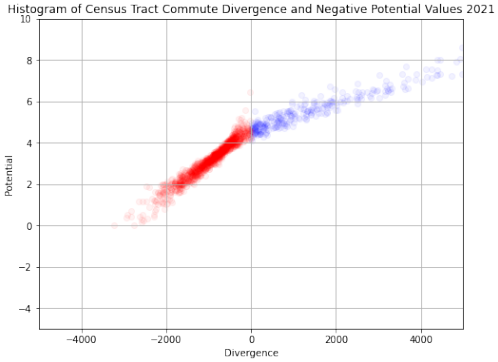
(a) Negative Potentials of Census Tracts in Chicago Metro Area



(b) Divergences of Census Tracts in Chicago Metro Area

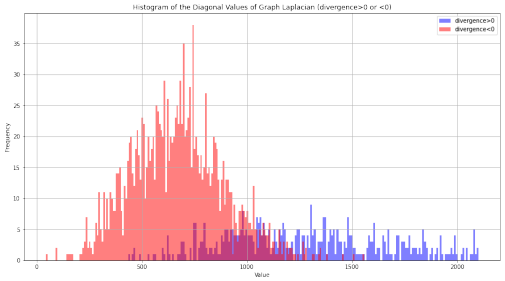


(c) Negative Potentials and Divergences of Census Tracts in Chicago Metro Area

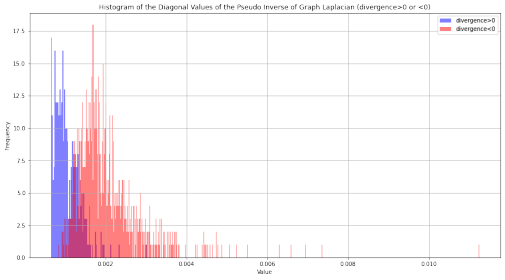


(d) Negative Potentials and Divergences of Census Tracts in Chicago Metro Area (Enlarged)

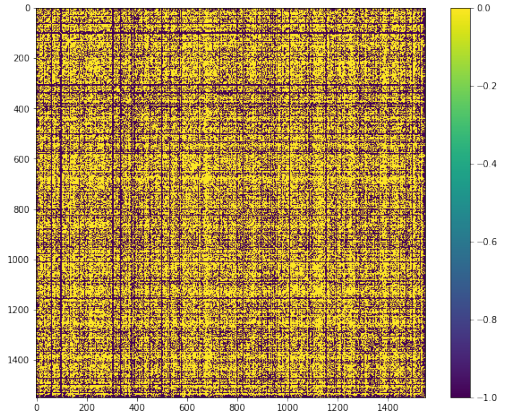
Figure 17: Negative Potential and Divergence in Chicago Metro Area



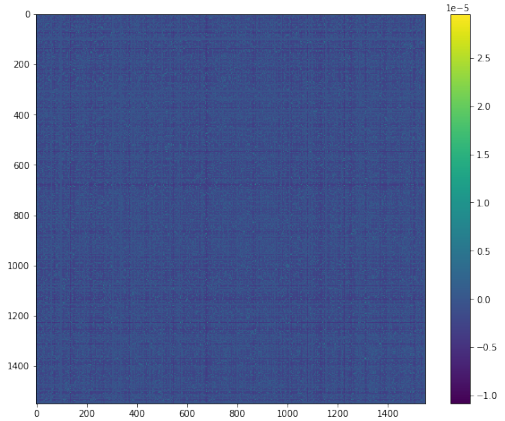
(a) Graph Laplacian Diagonal Values



(b) Graph Laplacian Pseudo Inverse Diagonal Values



(c) Graph Laplacian Non-Diagonal Values

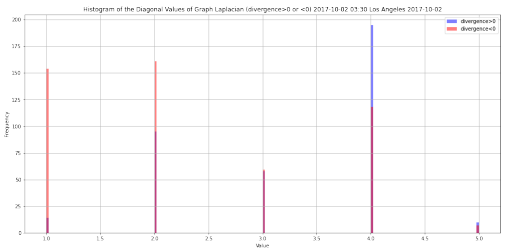


(d) Graph Laplacian Pseudo Inverse Non-Diagonal Values

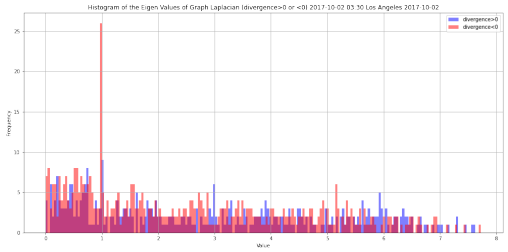
Figure 18: Negative Potential and Divergence in Chicago Metro Area

Appendix D

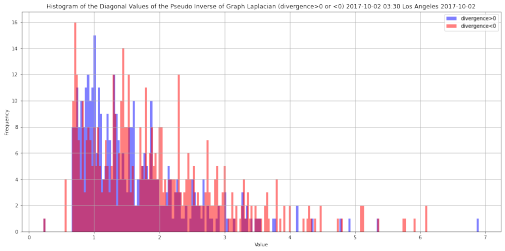
Figure 19: Road Network of Downtown LA



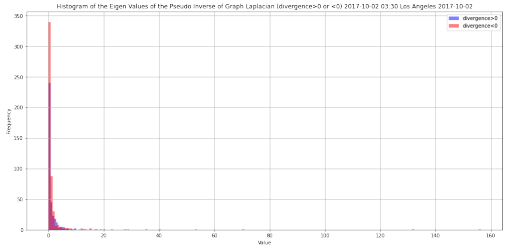
(a) Degree Distributions of Road Network of Downtown LA



(b) Eigenvalue Distribution of Graph Laplacian of Road Network of Downtown LA



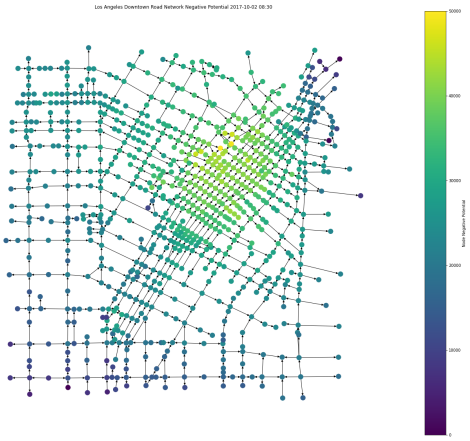
(c) Diagonal Values of Pseudo Inverse of Graph Laplacian of Road Network of Downtown LA



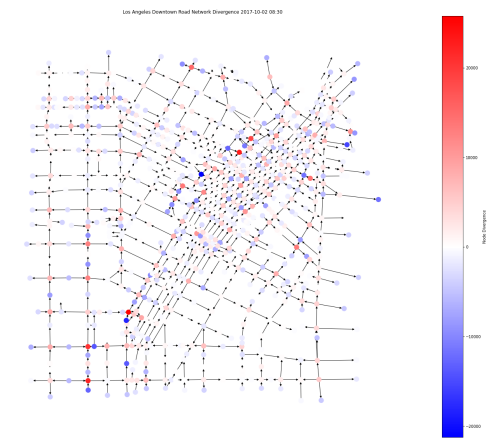
(d) Eigenvalue Distribution of Pseudo Inverse of Graph Laplacian of Road Network of Downtown LA

Figure 20: Properties of Graph Laplacian of Road Network of Downtown LA

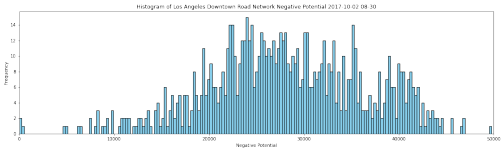
Appendix E



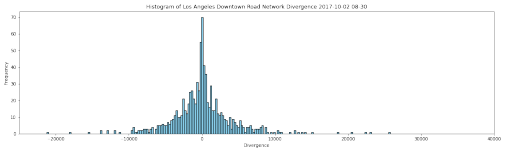
(a) Potential in Road Network of Downtown LA at 8:30



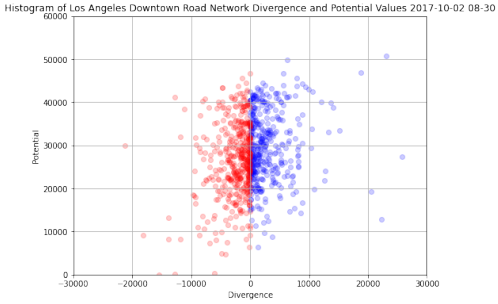
(b) Divergence in Road Network of Downtown LA at 8:30



(c) Potential Distribution in Road Network of Downtown LA at 8:30

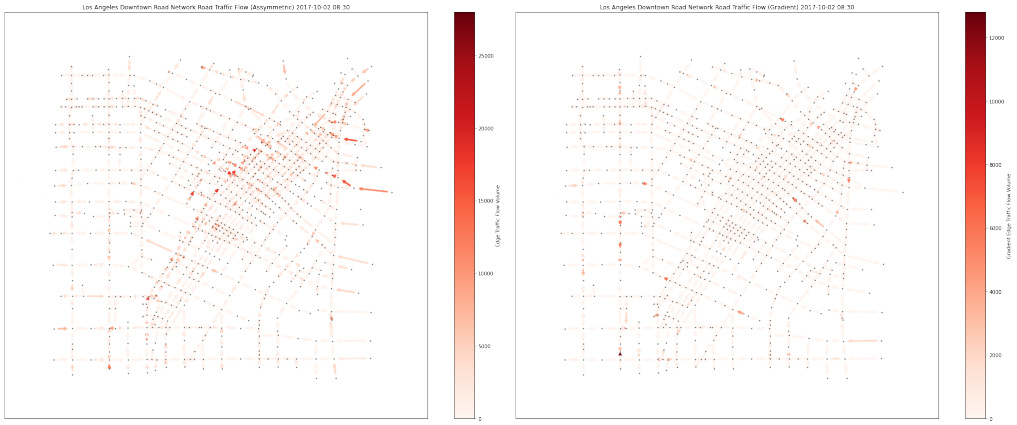


(d) Divergence Distribution in Road Network of Downtown LA at 8:30



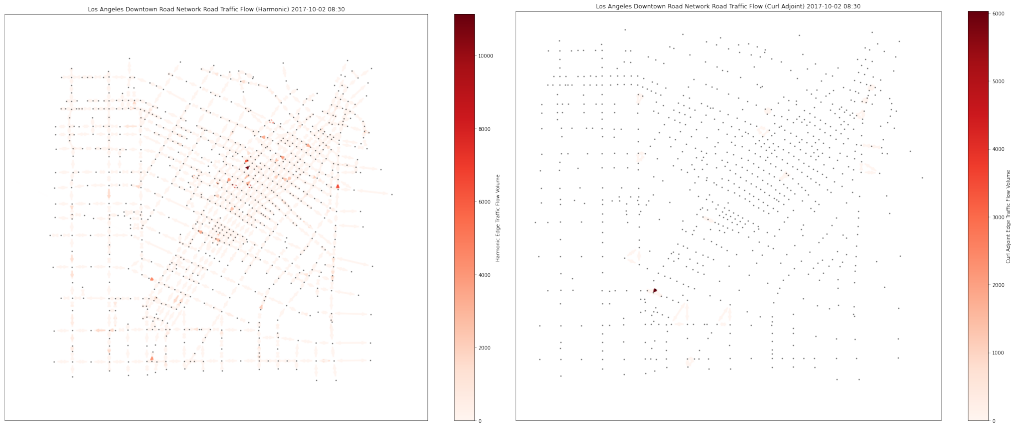
(e) Potential and Divergence in Road Network of Downtown LA at 8:30

Figure 21: Potential and Divergence in Road Network of Downtown LA at 8:30



(a) Edge Flow Raw Data in Road Network of Downtown LA at 8:30

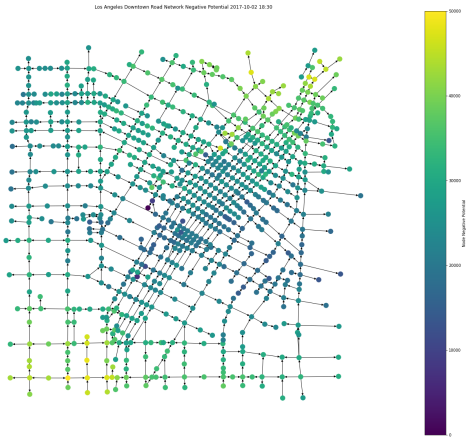
(b) Gradient Component in Edge Flow in Road Network of Downtown LA at 8:30



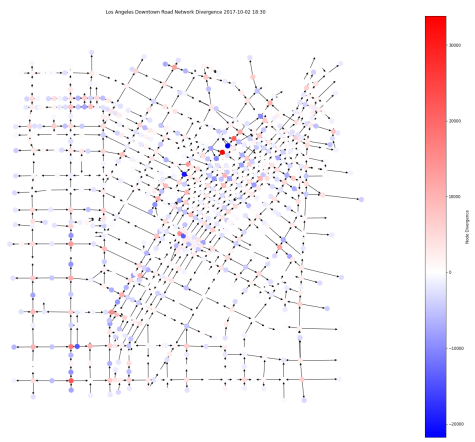
(c) Harmonic Component in Edge Flow in Road Network of Downtown LA at 8:30

(d) Curl Adjoint Component in Road Network of Downtown LA at 8:30

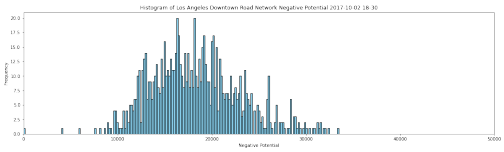
Figure 22: Hodge Decomposition in Road Network of Downtown LA at 8:30



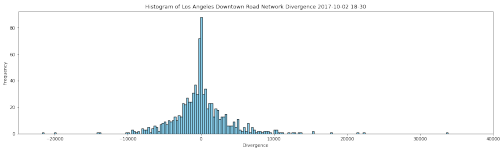
(a) Potential in Road Network of Downtown LA at 18:30



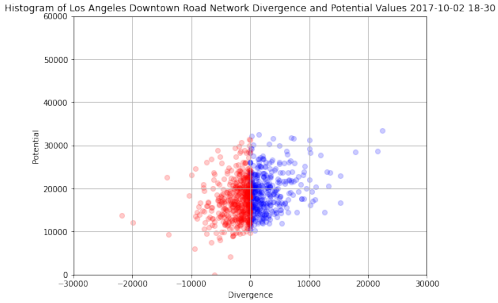
(b) Divergence in Road Network of Downtown LA at 18:30



(c) Potential Distribution in Road Network of Downtown LA at 18:30

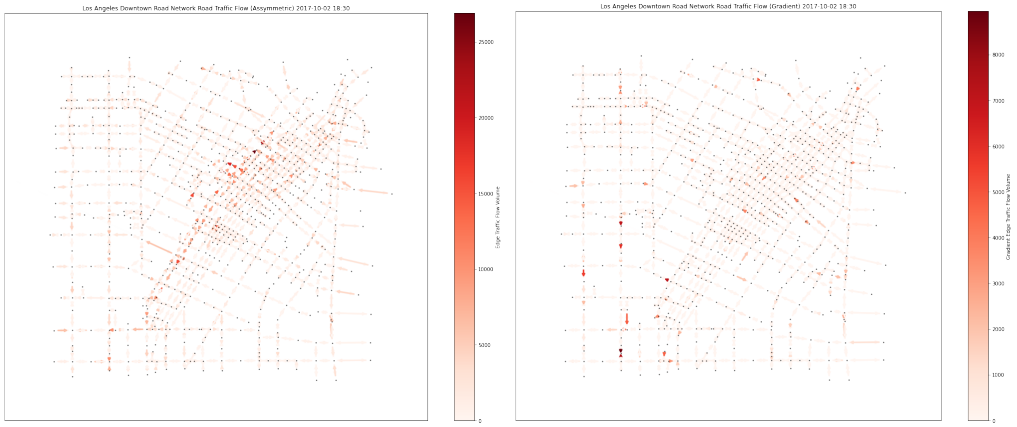


(d) Divergence Distribution in Road Network of Downtown LA at 18:30



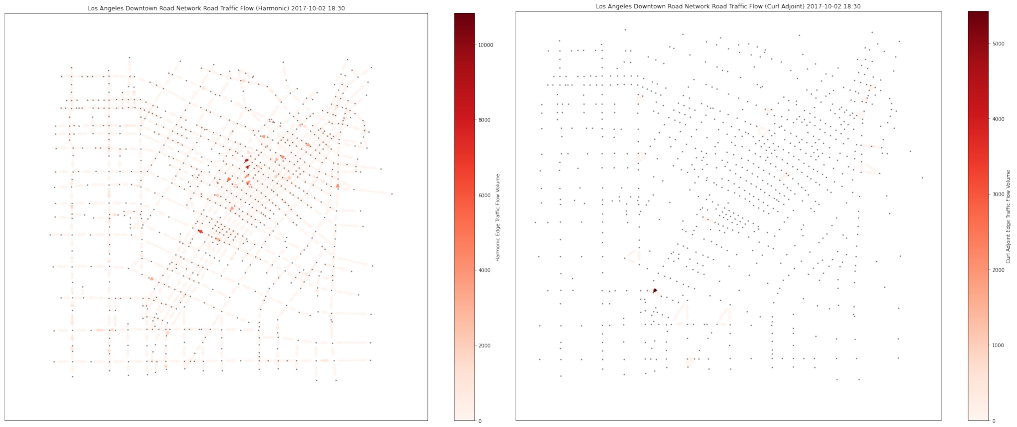
(e) Potential and Divergence in Road Network of Downtown LA at 18:30

Figure 23: Potential and Divergence in Road Network of Downtown LA at 18:30



(a) Edge Flow Raw Data in Road Network of Downtown LA at 18:30

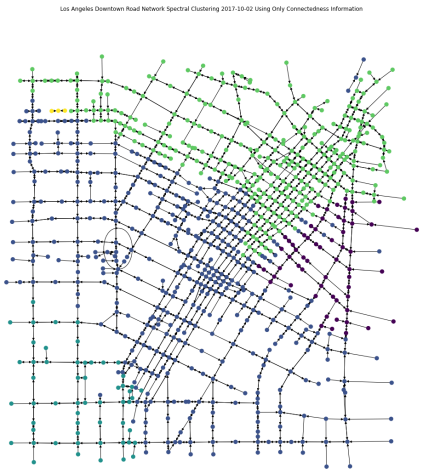
(b) Gradient Component in Edge Flow in Road Network of Downtown LA at 18:30



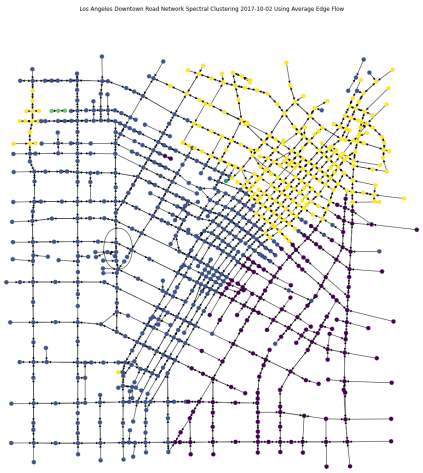
(c) Harmonic Component in Edge Flow in Road Network of Downtown LA at 18:30

(d) Curl Adjoint Component in Road Network of Downtown LA at 18:30

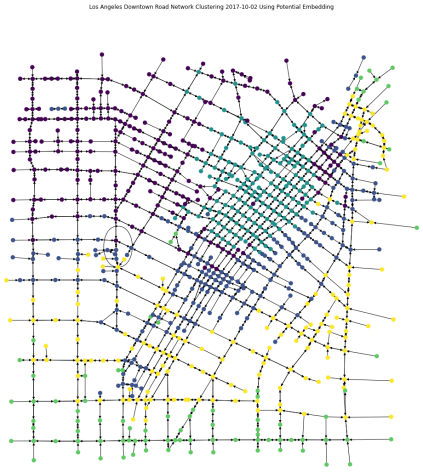
Figure 24: Hodge Decomposition in Road Network of Downtown LA at 18:30



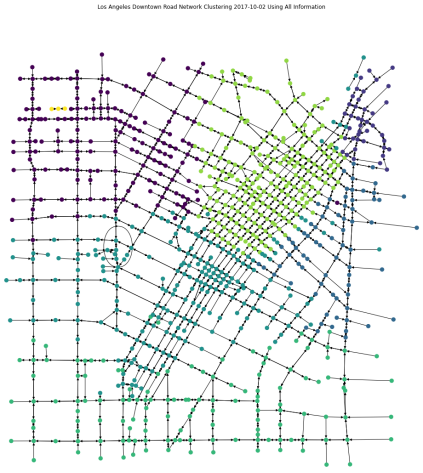
(a) Clustering of Road Network of Downtown LA Using Only Network Structure



(b) Clustering of Road Network of Downtown LA Using Only Mean Flow Volume on Edges



(c) Clustering of Road Network of Downtown LA Using Only Potential Embedding



(d) Clustering of Road Network of Downtown LA Using All Information

Figure 25: Clustering of Road Network of Downtown LA