Investigating machine learning for simulating urban transport patterns: A comparison with traditional macro-models

Citation for the original published paper (version of record):
Parishwad, O., Jiang, S., Gao, K. (2023). Investigating machine learning for simulating urban transport patterns: A comparison with
traditional macro-models. Multimodal Transportation, 2(3).
http://dx.doi.org/10.1016/j.multra.2023.100085

Multimodal Transportation 2 (2023) 100085

Contents lists available at ScienceDirect

Multimodal Transportation

journal homepage: www.elsevier.com/locate/multra

Full Length Article

Investigating machine learning for simulating urban transport patterns: A comparison with traditional macro-models

Omkar Parishwad[a,*], Sida Jiang[b], Kun Gao[a]

[a] Chalmers Institute of Technology, Architecture and Civil Engineering, Gothenburg, 41296, Sweden
[b] WSP,Transport Advisory, Globen, Stockholm, 12188, Sweden

ARTICLE INFO

Keywords:
Intelligent transportation systems
Passenger flow prediction
Spatial interaction model
Traffic simulation
Sensitivity analysis
Transport planning
Norway

ABSTRACT

Predicting passenger flow within a city is crucial for intelligent transportation management systems, especially in the context of urban development, post-pandemic policy changes, and infrastructure improvements. Traditional macro models have limitations in accurately capturing the complex structure of real traffic flows, and recent advancements in machine learning offer promising approaches for improving transportation simulations. This research aims to compare the effectiveness of traditional simulation models with a selective machine learning (ML) model for traffic flow prediction in Oslo, Norway. Sensitivity and scenario analyses are conducted to examine the models' parameters and derive the city's characteristics.

Results substantiate that the traditional Spatial Interaction model (SIM), although interpretable and requiring fewer parameters, has limitations in accurately capturing real flow structures and exhibits greater variability compared to the ML model. Statistical analyses support these findings and raise questions about the validity of the ML model's results over the SIM. The research highlights the potential of ML models to identify trends in passenger flows and simulate traffic flows in different scenarios related to city development. Overall, the research presents a decision support system for planners and policymakers to predict traffic flow accurately and efficiently. It highlights the benefits and drawbacks of both the traditional SIM and ML models, contributing to the ongoing discussion of the role of machine learning in transportation modeling.

1. Introduction

The Land Use Transport Interaction (LUTI) models have been theorized upon correlating factors by transport and city planners to explain the existing flows within the city (Wegener, 2004). Various endogenous phenomena metrics, such as measurement of workplace accessibility (Waddell, 2002), long-term implications of policy decisions, transportation infrastructure, and trip production activity distribution, real-estate and housing supply, goods and passenger transport distribution, incorporated into urban transportation or public services, were based on the classic four-stage sequential approaches, and have been critiqued on occasion for being out of date (Ma et al., 2021). This research focuses on accessing the simulations for a transport model for such use cases against traditional methods over modern machine learning algorithms.

* Corresponding author at: Chalmers University of Technology, Sweden.
E-mail address: omkarp@chalmers.se (O. Parishwad).

https://doi.org/10.1016/j.multra.2023.100085
Received 20 September 2022; Received in revised form 10 March 2023; Accepted 12 March 2023

1.1. Literature review

The research for the traditional LUTI models can be grouped into three generations (Cordera et al., 2017), theorized earliest during the 1960-70s. Models based on Wilson's statistical mechanic theory (Wilson, 2021) and Andrew's economic base theory, leading to Lowry's interaction model, are grouped under Spatial Interaction Models (Batty, 1976). Mathematical programming models under principles of agent behavior simulation, Alonso's theory of maximization of aggregated rents, and Input/Output matrices modeling based on Leontief and simulating regional economy, leading to models such as MEPLAN by Echenique to define the first generation of models (Wegener, 2021). The second and third-generation models were proposed during the 1980-90s based on the Random Utility Theory developed by McFadden and the dynamic micro-simulation models (Iacono et al., 2008) developed after the 1990s. Santiago's Land Use Model (MUSSA), developed by Martdnez simulating real estate markets, UrbanSIM developed by Waddell et al. (2007), followed by cellular automata models (Liu, 2008) simulating land use and changing upon specific rules of behavior. These proposed theories have evolved and are still being researched for relevant use cases (Zhong et al., 2022).

In addition to traditional transport models, relevant user equilibrium models, such as the General Stochastic Ridesharing User Equilibrium Problem with Elastic Demand (GSREU) (Ma et al., 2022), should be mentioned. The GSREU model accounts for the stochasticity of ridesharing demand and the elasticity of travel time, making it a more realistic representation of transportation behavior. Other user equilibrium models like the traditional Wardrop User Equilibrium (UE) (Zhang et al., 2011) or the stochastic network equilibrium have also been considered for incorporation. Other relevant urban transportation network models like the System Optimal models (Patil and Ukkusuri, 2007) optimize network performance by considering the whole network rather than optimizing individual trips. This approach can lead to better network performance overall but may result in an inequitable distribution of travel time across users. Therefore, both user equilibrium and system optimal models have advantages and disadvantages, and the choice of model should depend on the specific use case.

Another relevant contribution to transport macro models predicting travel patterns is space-time prisms (STPs), capable of analyzing human mobility patterns by integrating space and time dimensions. The traditional STPs are usually used to depict the movement trajectories of a single mode, such as walking or driving; they cannot capture the complex interactions between different modes in a multimodal transport system. However, the space-time prisms in multimodal supernetworks (STPMSNs) can comprehensively analyze the spatiotemporal accessibility and equality significance of different modes of transportation. The proposed STPMSN method has been applied to a case study of multimodal transportation in a metropolitan area (Qin and Liao, 2022). The results demonstrate its effectiveness in evaluating the accessibility and equality of different modes of transportation and identifying potential areas for improvement in the transport system.

This research is based on the traditional four-step transport model (McNally, 2007) approach. Growth-factor modeling as a Cordon model, which derives variable multipliers based on the relative trip change, can derive production and attraction for the trip generation stage. Cross-classification models, which group households into homogeneous groups for multiple-class analysis, produced better study results, while regression analysis produced the best results for trip generation, specifically at the activity end. The trip distribution was primarily influenced by the sound principles of Gravity modeling theory, and Hyman's method was used for deterrence calibration. Generalized costs were considered based on data availability, with distance-based costs as a baseline.

While traditional models have clear advantages, such as being easy to interpret and requiring few parameters, they need to capture actual flows' structure accurately and with more significant variability. This research attempts to decipher the black box for ML-derived simulation results and establishes ML algorithms' pre-eminence over the traditional SIM for simulation results. However, the validity of results for the ML model over the SIM remains a question. Sensitivity analysis and scenario analysis have been conducted to derive city character and decipher the validity of results.

2. Evaluation framework

The methodology presented in this study is generic and adaptable for any city worldwide, with customizations for the datasets. However, it is tailored for travel behavior analysis for the Oslo municipality. The study can be conducted in stages as defined ahead.

2.1. Defining the city system

The city subdivisions are defined based on data availability for deriving the macro simulation model of traffic flows (Dennett and Wilson, 2013). The subdivisions demarcate sectors for origin and destination to analyze trip patterns, considering a generalized region for multiple points of origin or destination for travel pattern analysis. Explanatory model variables are derived for various socioeconomic, demographic, and land use representative fields and analyzed to formulate models for transport patterns, defining accurate trip flows or activity patterns for the predefined city subdivisions. The variables are feature engineered or excluded based on sensitivity analysis of the model for optimized correlation statistics.

2.2. Transport Macro model and predictions

The research builds upon Transport macro models to acquire predictions for variation in the model variables as a scenario. Two models are developed and compared for prediction: the Gravity theory-based Spatial Interaction Model (Pagliara and Wilson, 2010) and the Machine Learning-based Regressor Ensemble Model (Nguyen et al., 2018).

2.2.1. Gravity-based spatial interaction model

The SIM establishes an analogy between the movements of people and the attraction of physical objects using the universal gravity law. It measures the consequences of changes in the conditions for the existing spatial flows by weighting the constraints such as population, employment, income zones for attraction and distances, cost of travel, or topography for deterrence to define probabilities of trips between zones or points within a region. The derived variable coefficient values in the case of constrained models define the weightage of variables for individual city subdivisions. The unconstrained SIM is evolved through singly-constrained: Origin or Production-constrained (Eq. (1)) and Destination or Attraction-constrained (Eq. (2)), and doubly-constrained (Eq. (3)) mechanisms and further enhanced (Sen and Smith, 2012; Senior, 1979) for data or city relevant entropy maximization, distance decay functions and such.

The singly-constrained SIM predicts our scenarios for changes in our production constraints (such as population) (Eq. (1)), and attraction constraints (employment, salary, income, etc) (Eq. (2)). Here, T_{ij}: flow or trips between 'i' and 'j'; K: true scaling constant or the constant of proportionality; β: scaling parameter; α, γ: Origin (O_i) Destination (D_j) attractors' explicit power functions; c_{ij}: deterrence as a combination of travel cost, distance and time taken to transverse between i and j; P_i, P_j: population attractors between 'i' and 'j' same as for Origin (O_i), Destination (D_j).

$$T_{ij} = A_i O_i D_j \exp(-\beta c_{ij}), \text{ subject to } \sum_{j=1}^{m} T_{ij} = O_i, (\text{location specific } \alpha) \tag{1}$$

$$T_{ij} = O_i B_j D_j \exp(-\beta c_{ij}), \text{ subject to } \sum_{i=1}^{n} T_{ij} = D_j, (\text{location specific } \gamma) \tag{2}$$

The doubly-constrained SIM (Eq. (3)) are the real-world scenarios where both our power functions (α, γ) are location dependent. Our training data provides us with these proportionality constants for modeling our correlations for the deterrents (travel cost variations).

$$T_{ij} = A_i O_i B_j D_j \exp(-\beta c_{ij}), \text{ subject to } \sum_{j=1}^{m} T_{ij} = O_i, \text{ and } \sum_{i=1}^{m} T_{ij} = D_j, (\text{location specific } \alpha, \gamma) \tag{3}$$

A common source of failure of the Poisson regression for the SIM is that the data needs to satisfy the mean equals variance- criterion imposed by the Poisson distribution. This model might be a good choice as it explains the excessive zero values well but fails if the converged result were set to False. However, the effects of fitting singly and doubly-constrained spatial interaction models using the Poisson regression approach are not encouraging (Flowerdew and Lovett, 1988). We cannot add origin and destination-specific predictors to a doubly-constrained model.

2.2.2. Machine learning based regressor ensemble models (EML)

The non-traditional hybrid model can be trained to understand the relationship between variables and evolve a single optimal solution to a problem, producing better predictive models and a single robust output. The EML equation can be expressed as:

$$y = \sum_{k=1}^{K} w_k f_k(x) + b \tag{4}$$

where y is the dependent variable, x is the independent variable, f_k is the kth base learner, w_k is the weight for the kth base learner, K is the total number of base learners, and b is the bias term.

Randomized decision trees are the foundation for the decision tree, Light-GBM, and extra-trees regressors. Linear regression, XGBoost (boosting), Random Forest regressor (bagging) algorithms, ridge, snap-boosting, snap-decision, and snap-random forest are a few EML algorithms that can be further optimized using feature engineering and hyperparameter optimizations (HPO), resulting in an excellent fit for our data. While bagging aims to produce an EML model with less variance than its base learners, boosting and stacking aim to produce strong models with less bias than their base learners. Ridge is similar to Ordinary Least Squares but penalizes the size of the coefficients. Errors such as overfitting and heteroskedasticity can be mitigated efficiently for ensembles. Cloud technologies can also be used for their AI automated modeling abilities for our simulation needs (Opara et al., 2022). The variables feature engineered to optimize the goodness of fit for our transport model can be manipulated to simulate the scenarios for analysis.

For both SIM and the ML model, each variable affects the city subdivisions in a weighted manner. Thus defining the sensitivity refers to the uncertainties in the models (Goldstein et al., 2015). A simple yet powerful way to understand a machine learning model is by doing sensitivity analysis and scenario analysis, where we examine each feature's impact on the model's prediction and identify the most important parameters and scenarios driving our results.

2.3. Simulating tool: Interactive web application

The potential of the research for its simulating needs can be tested by developing a user-friendly interactive web application to initiate variable changes, analyze predictions through the trained model, visualize the output over a map, and generate data infographics. The application allows users to interact with the model inputs and explore the impacts of various scenarios for the model outputs. These insights would be helpful for city planners as a support system for decision-making for development projects.

3. Case of Oslo, Norway

We follow up on the research methodology formulated in the earlier section, we delimit the study area as the Oslo municipality boundary, subdivided into 16 Urban divisions and the external Marka Forest, and further into 591 Grunnkrets (borough). The trip data was obtained from the Telia crowd insights (Company, 2021) dataset, while the variables for the city subdivisions were acquired from the official statistical online repository of Norway, Statistisk Sentralbyr (National Statistical repository).

Oslo is shaped by its natural setting, with 68% of the municipal area having publicly accessible green space surrounded by the Marka forest, which extends within and beyond the city. In 2019, the population of Oslo was 677,139, with the highest population density of 8600 people per square kilometer in its core areas. Oslo's population is almost 13% of Norway's population, with a growth rate of 1.40% for 2019.

3.1. Variables for model development

Due to the computational limitations and the desire to obtain meaningful insights, we choose to work with the urban divisions of Oslo municipality. The variables used to model the travel patterns were acquired from Statistikkbanken, Norway (Statistics Norway, 2022) and are described herewith.

(a) Population: The residential population within Oslo subdivisions was available for gender and age-specific counts. However, we found that the population correlation did not vary significantly for gender-specific or at the destination but did correlate highly for the age or independent age group of 19 to 69 years at the trip origin. Therefore, we considered only the independent (19-69 years) population counts at origin for all Urban Districts. Variation in population count would mean either demolition, newly proposed residential areas, or deaths during a pandemic or lockdown scenario.

(b) Employment: The number of jobs in all industries aggregated at the Urban District subdivision correlates to the destination incoming traffic counts. We assumed that the Urban structure of Oslo is such that the job count correlates with the land uses other than residential at the destination. Variation in employment could mean 'Work from Home' policy considerations or a proposal for new office premises.

(c) Household Income Level: The household income level is the average household gross income of anyone over 15. Interestingly, the zones with high-income groups are less populated and have low employment. Also, the wealthy stay away from the city area like the Marka forest and can travel longer distances. Variation in income level would mean a considerable migration due to some development, such as a proposed industry area or IT zone.

(d) Cost of Transport: The transport costs are directly proportional to the travel distance or travel time. However, since travel time varies heavily on the time of day, it is not fair to assign time variables to our travel. As we have zone-to-zone flows, we consider the centroid of the zone and identify the node nearest to its zone centroid. Then the shortest path was derived to every other nearest node from other zones using the Dijkstra's (Chao and Hongxia, 2010) shortest path algorithm. This distance between every coupled zone centroid defines the travel distance for deriving the transport cost. The derived shortest routes between urban Districts with trip counts are shown in Fig. 1(c). A variation in transport cost would predict the proposal of a new route, congestion for a zone or rise in fuel prices, or toll charges of some nature.

(e) Built Area: The built area comprising residential, commercial, industrial, and public-semi-public areas, constitute the urbanization level for a zone. Although the name depicts a relation to the area of the region or the proportion of Urbanization, that is not the case here. The trip generation is due to Urbanization which is accounted as the built area. A variation in Urbanization would mean new proposals for city development or demolitions.

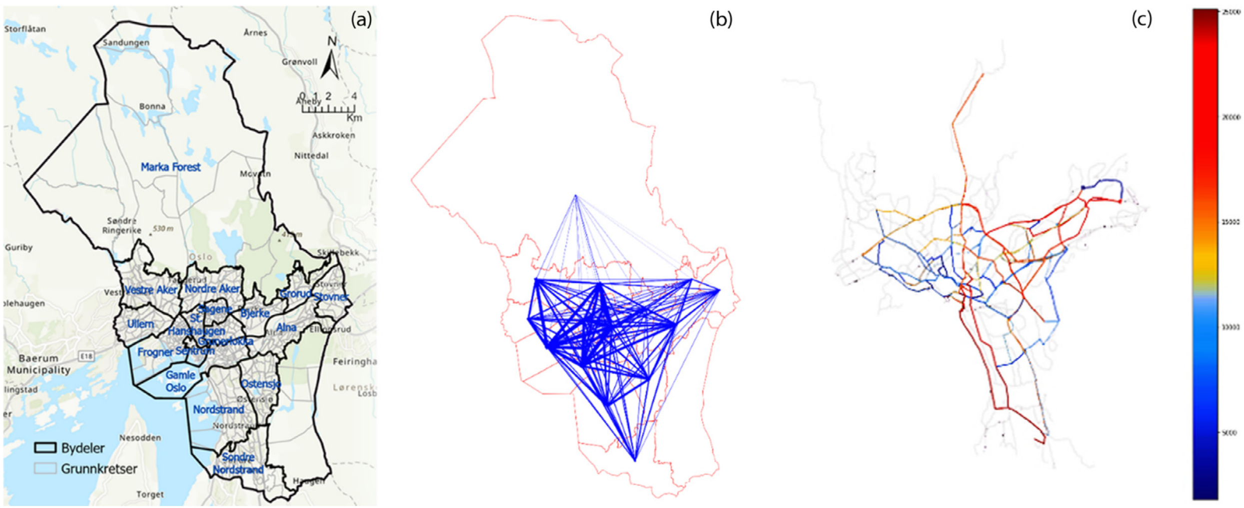

Fig. 1. Oslo Urban Districts (b) Trip counts (line thickness) (c) Shortest trip route with trip counts.

Variables such as literacy levels, student and foreign population counts, land use statistics, land value or housing prices were eliminated due to low sensitivity for the transport models. Transport network variables assigning accessibility, mode of transport, and journey time, to derive accurate travel costs were not incorporated in the model due to their absence in the acquired Telia mobile data (Company, 2021) due to GDPR (GDPR, 2016) constraints. This could have led to interpolating estimates, and a much better transport model, by incorporating the discrete choice models (Labbé et al., 1998).

The variables mentioned above were eventually used to develop the travel demand models for Oslo. These were chosen thru sensitivity analysis, reported in the later Section 4.2.1 of this research. The models were calibrated and validated using the available trip data. The calibrated model was then used to estimate the travel demand for different scenarios, including changes in population, employment, and urbanization levels on travel demand.

Overall, the case study of Oslo provides an excellent opportunity to analyze and model the travel demand for a medium-sized city with a diverse range of urbanization levels and a growing population. The study results can inform transportation planning and policy decisions in Oslo and other similar cities.

4. Transport modelling: Analyzing Oslo's travel patterns

The transport flows within Oslo city were modeled as behavioral patterns to forecast short-term simulations over a gravity-based Spatial Interaction Model (SIM) and a machine learning (ML) model. The SIM is a well-established, traditional transport planning model that has been used extensively to estimate travel flows. On the other hand, the ML model is a more recent approach that uses advanced algorithms to learn patterns from data and make predictions. By comparing the results of these two models, we can evaluate their strengths and weaknesses in capturing the complexities of travel behavior and identify areas for improvement.

For this study, we have adapted transport flows for all weekdays using data from the Telia crowd insights dataset. The dataset procured, aggregates flows to daily counts and between city subdivisions. A significant limitation of the gravity-based Spatial Interaction Model (SIM) is that it cannot account for intra-divisional trips, up to 40.5% of all trips. Additionally, the unavailability of OD geocoordinates for trips as they are aggregated at zone level also poses a challenge for the model. Nonetheless, predictions for the model for various scenarios are acquired by tuning the constants.

The following sections describe the SIM and the enhancements we made to incorporate independent variables. We then present the ML model and compare the results of these two models under different scenarios, using scenario analysis, and examine the sensitivity of the ML model to changes in the input variables.

4.1. The gravity-based spatial interaction model (SIM)

The SIM used in this research is based on a family (Dennett and Wilson, 2013; Pooler, 1994; Wilson, 1971) of similar incremental models developed over the years. The model starts with data exploration, and as the data is heavily right-skewed, it needs to be transformed to acquire a correlation and get closer to fitting a model estimate using a straight line. Also known as the Poisson distribution (Flowerdew and Lovett, 1988), the method states that the model's independent variable- trip counts, corresponds to the expected value, and the equation can be linearized. The leading theory behind the Poisson regression model is that the flows that spatial interaction models deal with (such as migration or commuting flows) relate to non-negative integer counts.

For the Oslo case study, the Production-constrained SIM is re-specified as a Poisson regression model by taking logs of the right-hand side of the equation and assuming that these are logarithmically linked to the Poisson distributed mean (Fig. 2).

$$\lambda_{ij} = \exp(\alpha_i + \gamma \log_e W_j - \beta \log_e d_{ij}) \tag{5}$$

where W_j is the destination employment, α_i is the equivalent of the vector of balancing factors A_i, but in regression or log-linear model terminology, it is a dummy variable or works as fixed effects. So, μ_i is modeled as a categorical predictor, and therefore in the Poisson regression model (Eq. 5), we do not use the numeric values of O_i; we can use a categorical identifier for the origin.

The Destination or Attraction-constrained Model produces results for variation in the population. For Oslo data, the correlation for the Origin-constrained Model is better than the Attraction-constrained Model. While the Doubly-constrained Model iteratively arrives at values for A_i and B_j by initially setting each to equal 1 and then calculating each in turn until the difference between each value is small enough not to matter as A_i relies on knowing B_j and the calculation of Bj relies on knowing A_i.

4.1.1. Enhancing the spatial interaction model

The SIM assumes that the distance decay parameter follows a negative power law (Taylor, 1971), but this may not always be the case. The negative exponential function ($e^{-\beta d_{ij}}$), is another standard function of distance used for the inverse power law. While the inverse power function has a far more rapid distance decay effect than the negative exponential function, in real life, if the observed interactions drop off very rapidly with distance, they might be more likely to follow an inverse power law. On the other hand, if the effect of distance is less severe, the negative exponential function might be more appropriate.

The SIM is limited to making predictions while keeping the total number of trips constant. When some or all the parameters are varied, the variation in inter-subdivisional trips does not occur as expected. To overcome this limitation, we vary the constants for the parameters $-\alpha$ for population, γ for employment, and β for the cost of transport while constrained. In conclusion, the SIM is a useful model for predicting spatial interactions between regions based on employment and population parameters. Table 1 shows a better fit during the weekdays when the work trips are suitably high. The model is set to perform for all data despite trip purpose and interestingly does well. It could be due to a high correlation between land use and urbanization parameters which lead to a specific

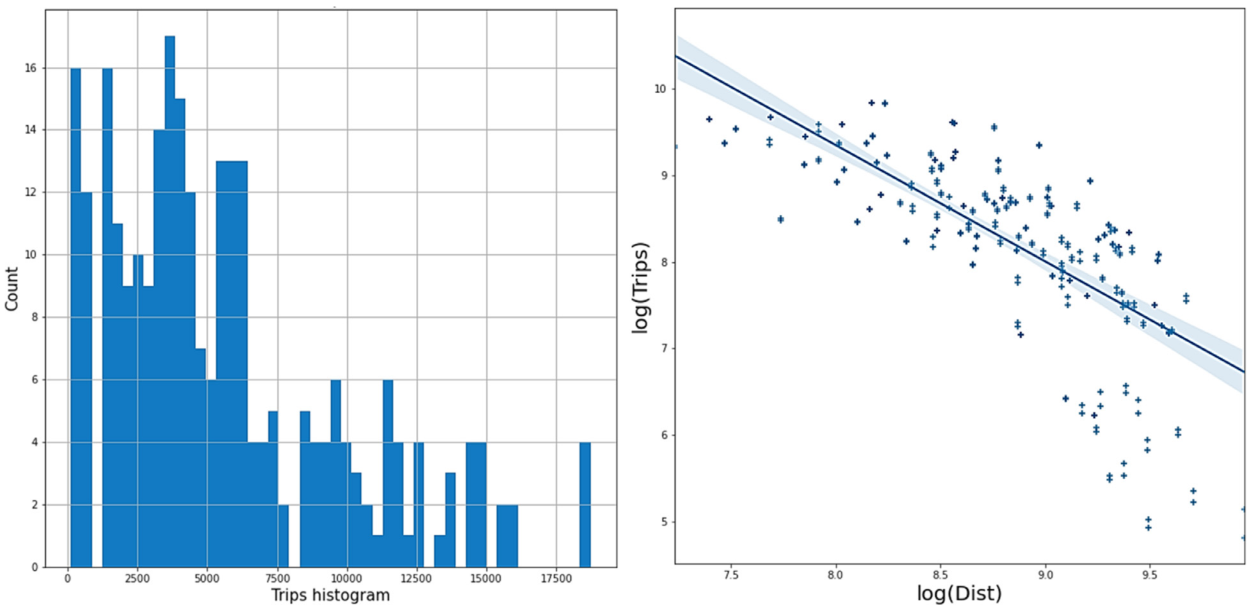

Fig. 2. Oslo trip flows histogram (left) and log transform correlation (right).

Table 1
Accuracy in terms of R^2 and Root Mean Square Error (RMSE) of the SIM.

	Orig-Constrained		Dest-Constrained		Doubly-Constrained		Distance Decay	
	RMSE	R^2(%)	RMSE	R^2(%)	RMSE	R^2(%)	RMSE	R^2(%)
AvgData	1106.19	83.46	1371.64	75.26	896.71	89.29	730.96	92.63
Monday	800.34	85.35	1016.12	76.9	625.79	91.09	596.29	91.82
Tuesday	1191.15	83.8	1479.11	75.64	964.82	89.45	786.29	92.74
Wednesday	1137.75	84.36	1444.73	75.4	920.33	89.86	743.69	93.14
Thursday	1269.6	84.28	1622	75.05	1033.43	89.72	834.91	93.04
Friday	1367.62	83.89	1747.98	74.48	1116.89	89.52	862.62	93.49
Saturday	1384.38	77.79	1589.37	71.72	1051.45	87.66	822.36	92.05
Sunday	896.53	75.7	956.63	72.67	676.06	86.15	599.93	88.76

trip purpose. By understanding the model's limitations and considering appropriate distance decay functions, we can enhance the model's predictive capabilities. While the model is limited in its ability to predict trip variations while keeping the total number of trips constant, varying the constants for the parameters can provide a potential solution.

Also, Table 2 provides insights into the push and pull constraints at the origin and destination, respectively. We can see that the push constraints at the origin accentuate traffic patterns significantly. Logically, the distance parameter provides the highest correlation, as expected in the SIM based on the work-trip theory. However, it is worth noting that the SIM is limited to employment and population parameters, identifying it as the work-home trips model. While the model fits better during weekdays when work trips are high, it performs well for all trip purposes. This could be due to the high correlation between land use and urbanization parameters, which can lead to a specific trip purpose.

Further deliberations of SIM for building comprehensive urban models (Oshan, 2021; Wegener, 2021) could not be adapted for this study due to the limitation of data required to develop them. Given the availability of data for basic and non-basic (manufacturing and public service employment), and demand for service and home-to-retail flows, we could have analyzed extended scenarios for employment, work trips, residential distribution, and retail demand for the city subdivisions.

4.2. *The machine learning model (ML)*

The Extreme Gradient Boosting (XGBoost) regressor algorithm is used over the Oslo trip dataset to predict transportation patterns in Oslo. The XGBoost algorithm is known for its ability to provide accurate predictions through second-order Taylor expansion using loss functions' derivatives, regularization parameters, and feature engineering. This algorithm is particularly efficient because it sorts traffic data by feature values before starting calculations and enabling parallel computing on feature enumerations.

The XGBoost algorithm performs wavelet decomposition and the threshold method to transform the data. The feature engineering approach is then used to derive variables that simplify the model after data transformations and improve accuracy. Unnecessary dataset features are removed from training to avoid making the model heavier and decreasing overall performance.

Table 2
SIM generated push (origin) or pull (destination) coefficients for each city subdivision.

City Subdivisions	Population	Employment	Distance			
	Orig. Const.	Dest.Const.	Doubly Const.		Dist.Decay	
	Origin	Destination	Origin	Destination	Origin	Destination
Alna	15.23	20.46	20.81	0.00	10.13	0.00
Bjerke	14.71	19.80	20.11	-0.58	9.51	-0.76
Frogner	15.39	20.47	20.78	0.14	10.32	0.00
Gamle Oslo	15.06	20.20	20.48	-0.30	9.94	-0.26
Grorud	14.56	19.77	20.08	-0.67	9.54	-0.63
Grünerløkka	14.95	20.01	20.34	-0.38	9.88	-0.37
Marka	12.75	18.13	18.40	-2.39	8.30	-1.85
Nordre Aker	15.09	20.13	20.53	-0.17	10.01	-0.25
Nordstrand	15.00	20.17	20.45	-0.27	9.85	-0.38
Østensjø	14.89	20.01	20.28	-0.41	9.71	-0.56
Sagene	14.49	19.58	19.82	-0.91	9.47	-0.77
Sentrum	15.29	20.45	20.73	-0.07	10.16	-0.03
Søndre Nordstrand	14.97	20.25	20.54	-0.22	10.37	0.21
St. Hanshaugen	15.18	20.27	20.62	-0.11	10.11	-0.13
Stovner	14.73	19.93	20.27	-0.47	9.94	-0.33
Ullern	14.78	19.82	19.98	-0.61	9.64	-0.77
Vestre Aker	14.99	20.24	20.54	-0.19	10.01	-0.24
$log Dempl$	0.33					
$log O Pop$		0.01				
$log Distance$	-1.20	-1.41	-1.40			
$Distance$						-0.0002

Table 3
ML Model Performance Metrics.

Model	RMSE	R-sq(%)	K-fold CV
All_Data	563.75	99.64	0.98
Monday	770.27	92.81	0.80
Tuesday	918.18	94.70	0.74
Wednesday	1091.73	93.55	0.78
Thursday	1612.74	84.87	0.74
Friday	1739.09	72.94	0.79
Saturday	1622.79	85.21	0.79
Sunday	1209.94	81.92	0.54

The hyperparameters of the XGBoost algorithm, which are specific values or weights that determine an algorithm's learning process, are then optimized using the RandomizedSearchCV method. This method randomly selects samples to find a better parameter combination, including common, booster, and learning target parameters.

The XGB model produced encouraging results using the five variables and the categorical weekday parameter. ML models are known to produce a better fit for larger datasets. The individual correlation for weekdays could be more encouraging. The weakest model performance is for Sunday since the base parameters for the model explicate work-home trips.

It is, however, unrealistically accurate while considering all the data with weekdays as categorical datatype and the intra-trips.

The model performs relatively better for the same data and has obvious advantages over the Gravity-based SIM. The XGBoost model is known for its ability to incorporate more parameters easily than the SIM. This study also derives an understanding of the prediction capabilities using statistical interpretation studies (Table 3).

The XGBoost model's strength lies in its ability to incorporate more parameters and produce more accurate (Zheng et al., 2020) predictions than the Gravity-based spatial interaction model. However, the model's unrealistically high accuracy may lead to overfitting, and it may perform better for smaller datasets. Overall, the XGBoost model provides encouraging results and demonstrates the potential of machine learning algorithms in predicting transportation patterns in Oslo.

4.2.1. Sensitivity analysis

Sensitivity analysis was conducted to evaluate the performance of the XGBoost machine learning model for Oslo's traffic patterns. It aimed to identify how changes in the model's input parameters affected its performance. This information is critical to understanding the model's limitations and improving performance. The sensitivity analysis was conducted by varying the hyperparameters of the XGBoost algorithm and observing the changes in the model's output. It highlighted the importance of selecting appropriate hyperparameters for the model to achieve optimal performance.

The sensitivity analysis in Fig. 3 is for traffic flows from Sangene to Ullern, two urban districts with distinct characteristics. Sangene is predominantly residential with a high population density and medium to low employment, while Ullern is highly commercial with

Fig. 3. Sensitivity percent changes for specific parameters for trips from Sangene to Ullern.

high employment, high income, and a low residential population. The analysis confirmed that most trips from Sangene to Ullern are work-related, as expected. The results in Figure 10 show that transport costs are inversely proportional to trip count, while employment is directly proportional. Interestingly, the population of Sangene has a minimal impact on trip counts, and the level of urbanization in Ullern only affects trip counts when it exceeds 10%. The income parameter also inversely affects trip counts. The ML model accurately predicted these relationships, evident for the highly residential Sangene district to the highly commercial Ullern district.

SHAP interpretations (SHapley Additive exPlanations:) We interpret further features using the TreeSHAP method (Lundberg et al., 2018) for gradient boosting models based on cooperative game theory, which is an improvement over traditional feature importance and partial dependence plots. The SHAP method (Lundberg and Lee, 2017) includes feature importance, feature dependence plots, local explanations, and summary plots.

The Shapely value (Shapley, 1953) is used to decipher the tree models and is computed as:

$$\phi_j = \frac{1}{K}\sum_{k=1}^{K}(\hat{g}(x^m_{+j}) - \hat{g}(x^m_{-j})) \tag{6}$$

where $\hat{g}(x^m_{+j})$ is the prediction for x, with a random number of feature values.

The TreeSHAP interaction values can be estimated as:

$$\phi_{i,j} = \sum_{S\subseteq N_{i,j}} \frac{|S|!(M-|S|-2)!}{2(M-1)!}\delta_{ij}(S), \text{ where } i \neq j, \text{ and} \tag{7}$$

$$\delta_{ij}(S) = f_x(S\cup i,j) - f_x(S\cup i) - [f_x(S\cup j) \pm f_x(S)] \tag{8}$$

with M being the number of features and S representing all feature subsets. The SHAP interaction value between each feature (i,j) is split equally, so the total interaction effect is $(\phi_{i,j} + \phi_{j,i})$.

The global (Fig. 4.2.1) and individual (Fig. 6) feature importance were visualized using summary and force plots. The global feature importance summary chart in Fig. 4 shows that 'Distance' has the most significant impact on prediction compared to other features. However, it combines prevalence and magnitude, which may not show the rare high-magnitude effects. The feature importance for individual prediction in Fig. 5(b) shows how the features are ranked by their effect on prediction and also depicts the impact distribution of each attribute. Here, 'Distance' has the most significant impact on predicting the number of trips, while "Income at destination" has the least impact. These force plots demonstrate how each feature contributes to the trip count predictions for the XGBoost model. Features that raise the prediction (to the right) are shown in red, while those that lower the prediction is shown in blue. There is an inverse relation between distance and direct proportion for the population. The high number of trips for 'Urbanization' for zones (Alna, Marka, Norde Aker, and Grunerlökka), while the high trip counts for the 'Sangene' zone can be visualized as a generic relation to the model for the data analyzed.

In conclusion, the SHAP method considerably improves model interpretability and allows for a better understanding of the relationship between feature variables and prediction results. It provides a comprehensive set of tools, including feature importance,

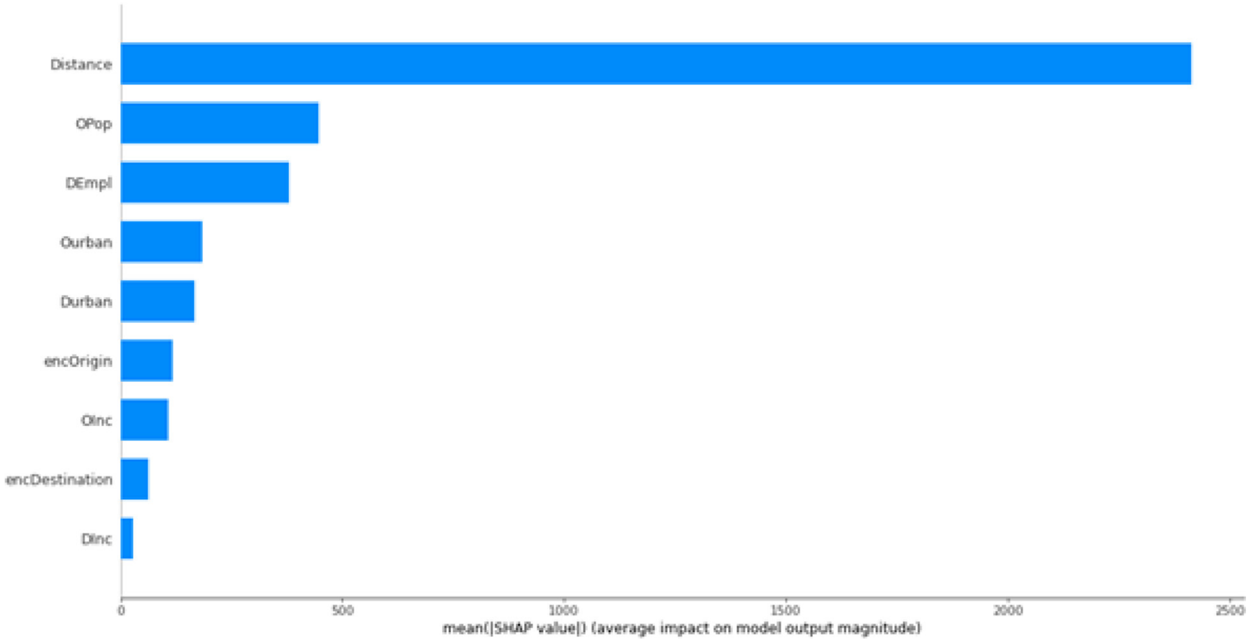

Fig. 4. Global feature importance for the independent parameters of our model.

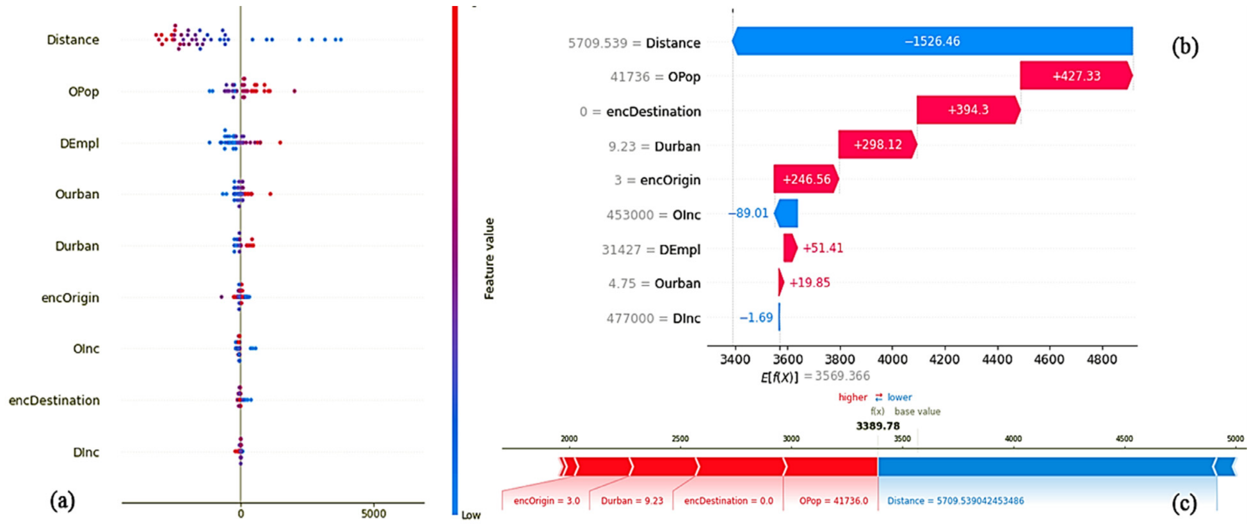

Fig. 5. SHAP summary plots (a) SHAP summary plot of XGBoost model (b) Waterfall diagram for individual prediction of SHAP values (c) First prediction by model using TreeSHAP.

feature dependence plots, local explanations, and summary plots, to enable the practical interpretation of complex machine learning algorithms.

5. Scenario analysis: Comparing the SIM and ML model predictions

Various characteristics of the Urban Districts and underlying assumptions regarding the purpose of the trip can explain predictions for trips between different districts in Oslo. However, the accuracy of these predictions is highly dependent on the relevant parameters selected for the model. This gives us a glimpse into the black box mechanism that processes these predictions in the backend. However, a scenario analysis is necessary to understand the parameters and corresponding statistics.

To conduct the scenario analysis, we used trip data for Oslo without the Intra-subdivisional trips to derive results for the ML algorithm and compare the SI Model statistics. One of the key advantages of the ML model is that it incorporates additional parameters such as income and urbanization, which allows for more accurate predictions of the number of trips between Oslo's Urban Districts.

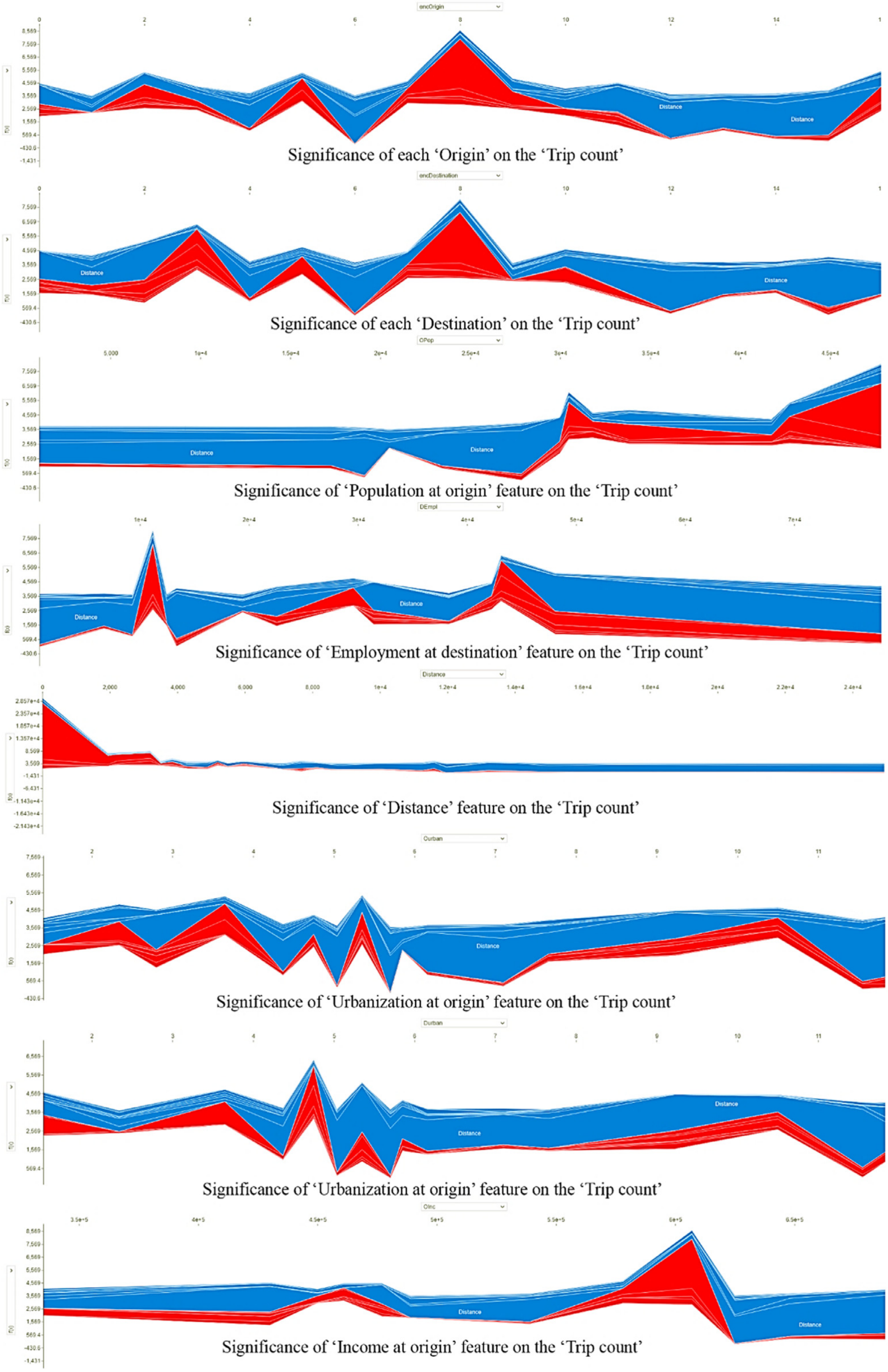

Fig. 6. Significance of independent features on the dependent 'Trip-count' feature.

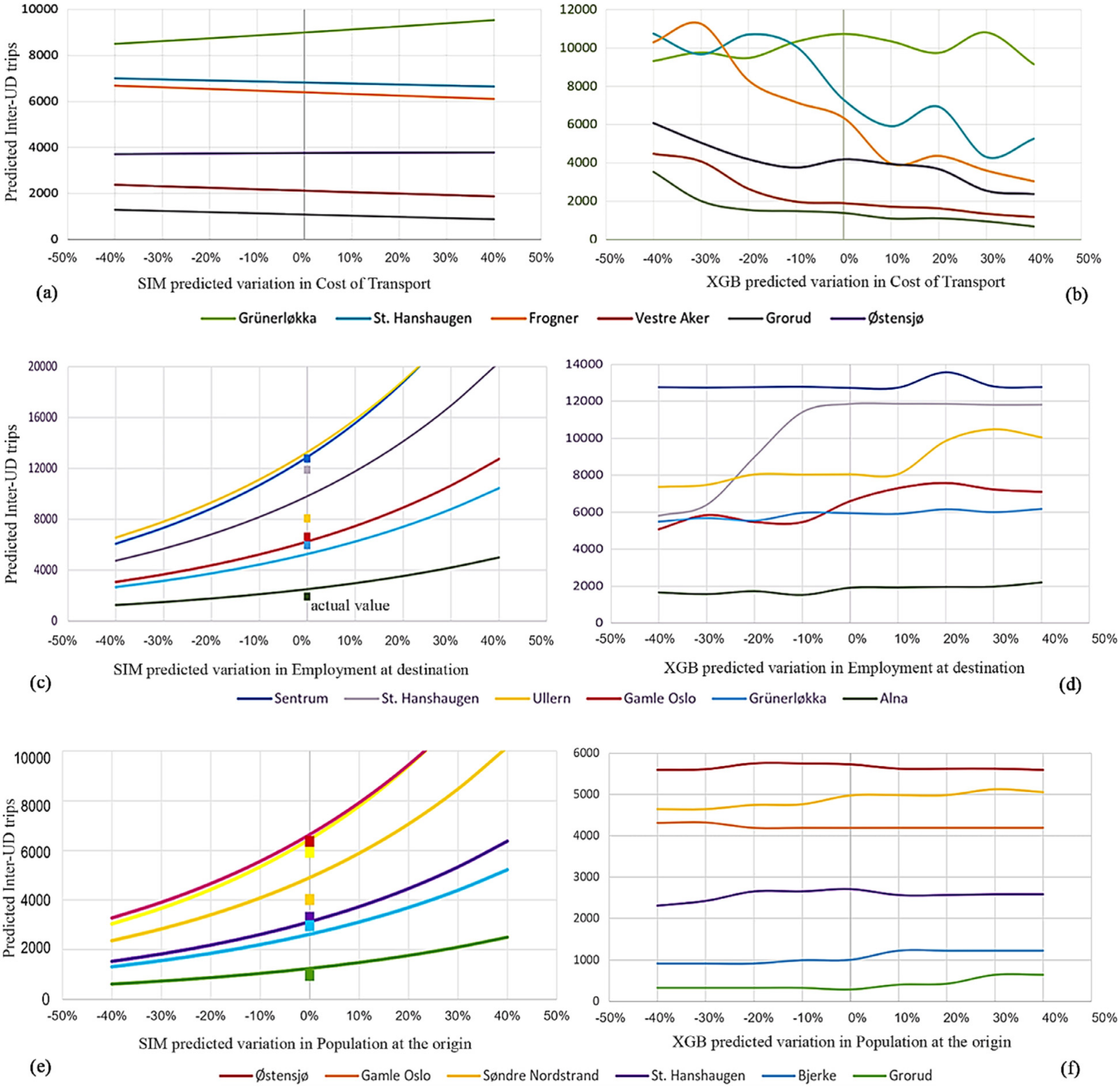

Fig. 7. Comparison of SIM (left) and ML-model (right) predictions (a,b) Outgoing trips from Gamle Oslo (c,d) Outgoing trips from Frogner (e,f) Incoming trips to Nordstrand.

In this section, we will focus on three comprehensive scenarios for analyzing trends and developing reasoning based on the accuracy of the results. These scenarios include uniform variations in a single parameter in a progression of 10% while keeping other parameters stable. The predictions are worked only for Wednesday as it has moderate yet high accuracy for both the transport models. By comparing the results of these scenarios between the two transport models, we can gain insights into which model performs better under different circumstances.

5.1. Scenario 1: Variation in cost of transport

Here, we explore the impact of a uniform variation in transport cost on outgoing trips from two subdivisions, Gamle Oslo and Grorud. This scenario is predicated on fuel price variations and congestion scenarios.

As expected, an increase in the cost of transport leads to a decrease in the number of overall trips. However, the decrease in the number of trips between the nearer districts is less significant than those further away, as seen in the ML predicted trips in Fig. 7(a,b).

Income and employment factors cause modest fluctuations in this pattern for the districts. Therefore, even with increased transport costs, trip counts may increase due to high income or employment, as other people utilize public transport.

The polynomial fluctuations in the predicted trips are also a result of the thresholds between the inter and intra-trips, which balance themselves. In Fig. 7 (a,b), we can see that only inter-trips from Gamle Oslo between subdivisions are present. The inter-trips from Gamle Oslo to the nearer Urban Districts with high employment potential, already having high inter-trips, have a significant variation in the predicted trips with variation in the transport cost. The Urban Districts further away, with even average employment, have only slight variation in the total inter-trips predicted by the models.

The SIM is based on the principle of constant total trips. Thus, we observe a negative slope for districts at a distance and a positive slope for nearer districts. Due to the doubly-constrained nature of this SIM, we have linear variations. However, comparing the predictions for the SIM and XGB models in Fig. 7, we see a significant similarity in the variations in the predicted inter-trips. The SIM has fewer parameters than the ML model (Cost of Transport, Population at Origin, and Employment at Destination), resulting in linear predictions. Additionally, the SIM discredits intra-trips due to its zero distance in its calculations.

5.2. Scenario 2: Variation in employment

In Scenario 2, we consider a uniform variation in employment across all the city subdivisions to predict the outgoing trips from specific subdivisions. This scenario could be relevant during a pandemic when work-from-home policies or relocation of company premises may impact employment in different areas. Figure 7 (c,d) shows the outgoing trips from Frogner with a variation in employment, while other parameters are constant.

We observe that the predictions from our two models differ significantly in this scenario. The ML model accounts for inter and intra-trips, and the variation in employment influences the population of the urban district in consideration. For example, we see a different behavior for St. Hanshaugen than in other urban districts. St. Hanshaugen is densely residential and has high employment, primarily catered to by the residential population within the urban district. So, when employment decreases, inter-trips decrease significantly in St. Hanshaugen. In contrast, other districts such as Sentrum, Ullern, Gamle Oslo, and Alna, which are situated closer and have a relatively lower population and higher employment, do not show significant variations with changes in employment.

On the other hand, the SIM assumes that even if the population is stable, higher employment leads to people attending more than one job. Thus, the population parameter is not prioritized in the model and predicted inter-trips vary exponentially with changes in employment. However, we also observe that the prediction accuracy of the SIM is relatively lower compared to the ML model, as shown in Fig. 7, where the square dots represent the actual trips, while the SIM predictions are consistent.

5.3. Scenario 3: Variation in population

Here, in Fig. 7 (e,f) the models predict incoming trips to Frogner and Nordstrand based on uniform variation in population across all city subdivisions. We find a high correlation between the ML predicted and actual trips, although the residual values for ML are higher. The SIM predicts that a higher population leads to exponentially higher trips, but there is a need for a more complex coupled spatial interaction SIM.

Overall, the matrices from the SIM and ML models for trip flows show roughly the same pattern, but ML performs better with low data availability, and SIM performs better at the Grunnkret level. The processing time for enhanced ML models is higher than SIM, and cloud-based applications like AWS, Azure, or IBM Watson are needed to handle huge matrices.

The research notes that the observations based on the predictions are limited since they uniformly varied population, employment, and cost of transport, which may not reflect specific development scenarios for a city. Hence a real-world scenario has been worked on over the developed visualization and simulation toolkit and explained next.

6. Application: Real-world scenarios

The real-world scenario analysis is a crucial step in understanding the parameters' influence on a city's transportation system. In this case, an interactive web application that includes parameters such as population, employment, income, urbanization or built area, and transport cost for all subdivisions in Oslo was developed (Parishwad, 2022). The dashboard is linked with cloud technology over IBM Watson in the backend, where a machine learning model processes the parameter variations as a real-world scenario for the city. The dashboard is designed to visualize the trip variation due to modifications made to the parameters over a node-node and the Oslo Road network, predicting the revised flow over the major streets. This enables the user to download the model analysis for further consideration.

In a specific scenario for Oslo, the dashboard predicts a scenario for a 5% decline in the population of St. Hanshaugen and a 10% growth in employment for Sentrum. This results in an increase in both the income and built area of Sentrum by 5%. Consequently, the number of trips in the city would vary, and the interactive dashboard provides graphs to visualize the change in the number of trips for further stages of four-stage modeling.

This real-world scenario analysis application allows decision-makers to understand the impact of changes in different parameters on the transportation system. It can help identify areas of concern, such as intersections that may encounter heavier traffic, and inform policy and design modifications to address them. Additionally, the interactive dashboard and machine learning model enable rapid and accurate analysis of scenarios, providing decision-makers with a powerful tool to improve transportation planning and management.

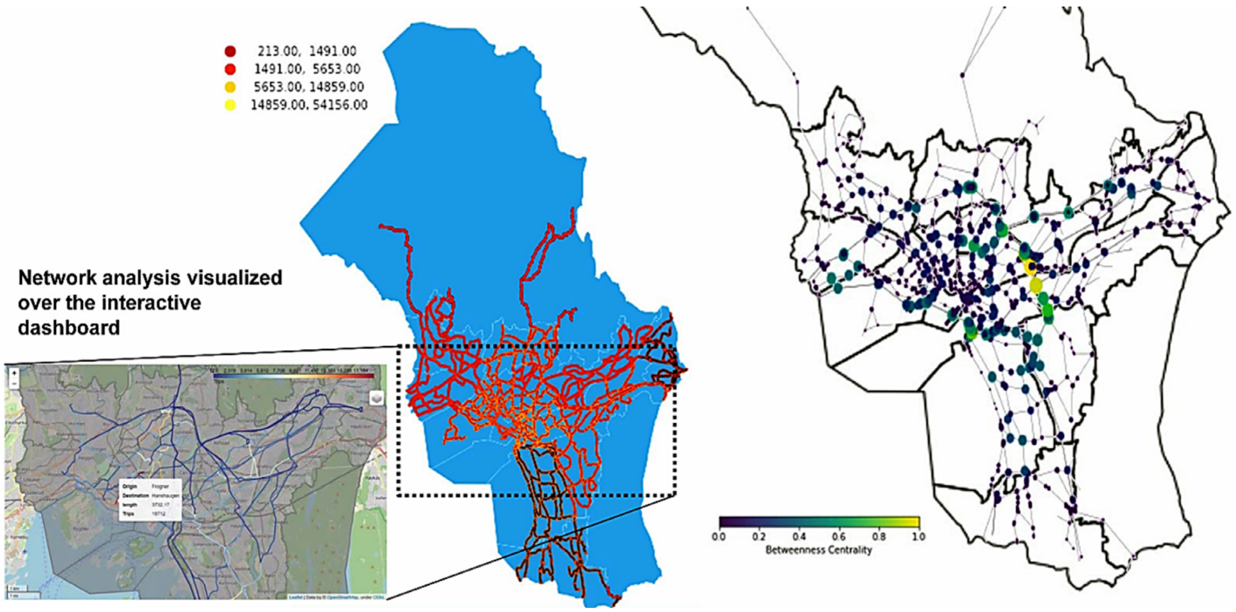

Fig. 8. Network analysis showcasing the betweenness centrality for Oslo Roads.

6.1. Network analysis and visualizations

In the above application, the network analysis also includes visualization techniques to understand the traffic flows over road segments and identify potential traffic congestion areas. Heat maps can be generated to show the traffic density on the road network, as shown in Fig. 8, which can help identify traffic congestion. The analyzed measures of centrality display the significant intersections that would affect the trips from and to every urban district's centroids. A weightage of trips is assigned to all the nodes, which defines their centrality scores. For example, the betweenness centrality scores the traffic flow, assuming the trips take the shortest path to interpret congestion. This can empower traffic management plans to reduce congestion and improve traffic flow. In addition, network graphs can represent the road network, with nodes representing intersections and edges representing the roads connecting them. This can help identify the shortest and most efficient routes between different locations within the city and areas where traffic may flow more slowly than expected.

Finally, simulation tools can model the traffic flow in real time, allowing traffic engineers and planners to test different traffic management scenarios and see how they would affect traffic flow throughout the city. This can help develop effective traffic management plans that can reduce congestion, improve traffic flow, and ensure the safety of all road users. Overall, the network analysis and visualization section provides vital insights into the traffic flow within the city, identifying areas where traffic congestion may be an issue and allowing for the development of effective traffic management plans to improve traffic flow and reduce congestion.

7. Inferences and conclusion

This research develops[1] a comprehensive transport model for trip predictions that can consider various real-world scenarios for their analytical interpretations. Similar studies (Tillema et al., 2006) investigating such comparisons do not consider logical hypothesis testing for their results, which is an integrated part of our research. Our analytical hypothesis testing and simulation-based approach using SIM and machine learning models have allowed for a comparative reconfirmation of the results, a unique aspect of our research. Our study highlights the data requirement for trips at the city level with considerable diversity, which is essential for any transport modeling.

Although this research is limited to the pre-COVID scenario of 2019, the post-COVID trends have revealed a significant variation in traffic flows in the city, which mandates a review of the public transport fleet and pricing mechanisms. Manufacturing, sales, and the food industry sectors need to return to their original working conditions. The rise in remote working has tipped trip purposes towards leisure activities, transforming the city into more people-centric (Venter et al., 2020). We foresee a shift towards the 15-minute city concept and a revision of the transit-oriented development for a dispersed city from a compact form.

This study obviates the importance of integrating spatial models for understanding human mobility. Concepts such as the Space-Time Prism (Qin and Liao, 2021; 2022) are fundamental for delineating individual behavior and constructing a sizeable multimodal

[1] Github repository: https://github.com/parishwadomkar/Investigating-machine-learning-for-simulating-urban-transport-patterns (accessed 06 June 2023).

transportation model. Integration of transport and spatial models (Zondag et al., 2015) in the development process could influence implementation decisions and affect the land use model structure, including accessibility, economic transition, employment location, and land price.

This research is limited to Oslo, but the application can be further tailored to provide predictions for multiple cities worldwide by acquiring real-time trip data through an API for training a model and analyzing real-time activity flows. API services that provide such worldwide real-time and historical routing data for cities are Otonomo, TomTom, and INRIX.

In conclusion, this research provides valuable insights into developing a comprehensive transport model for trip predictions that considers a wide range of real-world circumstances and their interpretations. Our study highlights the importance of integrating spatial models and the need for data requirements for trips at the city level. The post-COVID trends have revealed the necessity for an interactive platform to accommodate a wide range of real-world circumstances and their interpretations. Our study provides a solid foundation for future research in developing transport models that are more inclusive and comprehensive and that take into account real-world changes and trends.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Batty, M., 1976. Urban Modelling. Cambridge University Press Cambridge.

Chao, Y., Hongxia, W., 2010. Developed Dijkstra shortest path search algorithm and simulation. In: 2010 International Conference on Computer Design and Applications, Vol. 1. IEEE, pp. V1–116. doi:10.1109/ICCDA.2010.5541315. https://ieeexplore.ieee.org/document/5541315.

Company, T., 2021. Crowd insights for municipalities. Online webinar and training, https://coda.io/@data-insights/telia-webinars-and-training/crowd-insights-for-municipalities-18.

Cordera, R., Ibeas, Á., dell'Olio, L., Alonso, B., 2017. A brief history of land use–transport interaction models. In: Land Use–Transport Interaction Models. CRC Press, pp. 3–10.

Dennett, A., Wilson, A., 2013. A multilevel spatial interaction modelling framework for estimating interregional migration in europe. Environ. Plann. A 45 (6), 1491–1507. doi:10.1068/a45457.

Flowerdew, R., Lovett, A., 1988. Fitting constrained poisson regression models to interurban migration flows. Geogr. Anal. 20 (4), 297–307. doi:10.1111/j.1538-4632.1988.tb00184.x.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J. Comput. Graphical Stat. 24 (1), 44–65.

Iacono, M., Levinson, D., El-Geneidy, A., 2008. Models of transportation and land use change: a guide to the territory. J. Plann. Lit. 22 (4), 323–340.

Labbé, M., Laporte, G., Tanczos, K., Toint, P. (Eds.), 1998. Discrete Choice Models. Springer Berlin Heidelberg, Berlin, Heidelberg doi:10.1007/978-3-662-03514-6_9.

Liu, Y., 2008. Modelling Urban Development with Geographical Information Systems and Cellular Automata. CRC Press.

Lundberg, S. M., Erion, G. G., Lee, S.-I., 2018. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 https://arxiv.org/abs/1802.03888.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Ma, J., Meng, Q., Cheng, L., Liu, Z., 2022. General stochastic ridesharing user equilibrium problem with elastic demand. Transp. Res. Part B Methodol. 162, 162–194.

Ma, Q., Li, S., Zhang, H., Yuan, Y., Yang, L., 2021. Robust optimal predictive control for real-time bus regulation strategy with passenger demand uncertainties in urban rapid transit. Transp. Res. Part C Emerg.Technol. 127, 103086.

McNally, M.G., 2007. The four-step model. In: Hensher, D.A., Button, K.J. (Eds.), Handbook of Transport Modelling, Vol. 1. Emerald Group Publishing Limited, Bingley, pp. 35–53. doi:10.1108/9780857245670-003.

Nguyen, H., Kieu, L.-M., Wen, T., Cai, C., 2018. Deep learning methods in transportation domain: a review. IET Intell. Transp. Syst. 12 (9), 998–1004.

Opara, E., Wimmer, H., Rebman, C.M., 2022. Auto-ML cyber security data analysis using google, azure and IBM cloud platforms. In: 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET). IEEE, pp. 1–10.

Oshan, T.M., 2021. The spatial structure debate in spatial interaction modeling: 50 years on. Prog. Human Geogr. 45 (5), 925–950. doi:10.1177/0309132520968134.

Pagliara, F., Wilson, A., 2010. The state-of-the-art in building residential location models. In: Residential Location Choice: Models and Applications. Springer, pp. 1–20.

Parishwad, O., 2022. Developing transport interaction macromodels to simulate traffic patterns: case of Oslo, Norway Thesis. http://kth.diva-portal.org/smash/record.jsf?pid=diva2:1667625.

Patil, G.R., Ukkusuri, S.V., 2007. System-optimal stochastic transportation network design. Transp. Res. Record 2029 (1), 80–86.

Pooler, J., 1994. An extended family of spatial interaction models. Prog. Hum. Geogr. 18 (1), 17–39.

Qin, J., Liao, F., 2021. Space-time prism in multimodal supernetwork - Part 1: methodology. Commun. Transp. Res. 1, 100016. doi:10.1016/j.commtr.2021.100016.

Qin, J., Liao, F., 2022. Space-time prisms in multimodal supernetwork-Part 2: application for analyses of accessibility and equality. Commun. Transp. Res. 2, 100063.

Regulation (eu), 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679™from=EN Accessed on February 28, 2023.

Sen, A., Smith, T.E., 2012. Gravity Models of Spatial Interaction Behavior. Springer Science & Business Media.

Senior, M.L., 1979. From gravity modelling to entropy maximizing: a pedagogic guide. Prog. Hum. Geogr. 3 (2), 175–210.

Shapley, L., 1953. A value for n-person games. In: Kuhn, H., Tucker, A. (Eds.), Contributions to the Theory of Games II. In: Annals of Mathematics Studies, Vol. 28. Princeton University Press, pp. 307–317. https://doi.org/10.1515/9781400829156-012 Chapter 7.

Taylor, P.J., 1971. Distance transformation and distance decay functions. Geogr. Anal. 3 (3), 221–238. doi:10.1111/j.1538-4632.1971.tb00364.x.

Tillema, F., Van Zuilekom, K.M., Van Maarseveen, M.F., 2006. Comparison of neural networks and gravity models in trip distribution. Comput.-Aided Civil Infrastruct. Eng. 21 (2), 104–119. doi:10.1111/j.1467-8667.2005.00421.x.

Venter, Z.S., Barton, D.N., Gundersen, V., Figari, H., Nowell, M., 2020. Urban nature in a time of crisis: recreational use of green space increases during the Covid-19 outbreak in Oslo, Norway. Environ. Res. Lett. 15 (10), 104075. doi:10.1088/1748-9326/abb396.

Waddell, P., 2002. Urbansim: modeling urban development for land use, transportation, and environmental planning. J. Am. Plann. Assoc. 68 (3), 297–314.

Waddell, P., Ulfarsson, G.F., Franklin, J.P., Lobb, J., 2007. Incorporating land use in metropolitan transportation planning. Transp. Res. Part A Policy Pract. 41 (5), 382–410.

Wegener, M., 2004. Overview of land use transport models. In: Handbook of Transport Geography and Spatial Systems, Vol. 5. Emerald Group Publishing Limited, pp. 127–146.

Wegener, M., 2021. Land-use transport interaction models. Handb. Reg. Sci. 229–246.
Wilson, A., 2021. Entropy in urban and regional modelling. In: Handbook on Entropy, Complexity and Spatial Dynamics. Edward Elgar Publishing, pp. 20–31.
Wilson, A.G., 1971. A family of spatial interaction models, and associated developments. Environ. Plann. A 3 (1), 1–32.
Zhang, C., Chen, X., Sumalee, A., 2011. Robust wardrop's user equilibrium assignment under stochastic demand and supply: expected residual minimization approach. Transp. Res. Part B Methodol. 45 (3), 534–552.
Zheng, S., Zhang, S., Song, Y., Lin, Z., Wang, F., 2020. A noise-eliminated gradient boosting model for short-term traffic flow forecasting. In: 2020 8th International Conference on Digital Home (ICDH), pp. 166–171. doi:10.1109/icdh51081.2020.00036.
Zhong, M., Hunt, J.D., Abraham, J.E., Wang, W., Zhang, Y., Wang, R., 2022. Advances in integrated land use transport modeling. In: Advances in Transport Policy and Planning, Vol. 9. Elsevier, pp. 201–230.
Zondag, B., Helbich, M., Van Der Vlist, A., van Maarseveen, M., van Wee, B., 2015. Accessibility modeling and evaluation: the TIGRIS XL land-use and transport interaction model for the Netherlands. Comput. Environ. Urban Syst. 49, 115–125. doi:10.1016/j.compenvurbsys.2014.07.003.
Statistics Norway, 2022. https://www.ssb.no/en/statbank Accessed: May 5, 2022.