Predicting commuting flows using urban indicators and machine learning methods

Stefanie Helfenstein, Leila Paolini, Marius Simon Wrobel
Project 2 of Machine Learning (CS-433)

***Abstract*—The goal of this project is to study commuter flows between Swiss municipalities and to evaluate machine learning models for geographically predicting both the existence and magnitude of commuting flows using spatial, demographic, and socio-economic indicators. We compare two traditional mobility models, gravity and radiation, with three machine learning approaches: XGBoost, CatBoost, and a fully connected neural network. The task is decomposed into a binary classification problem for flow existence and a regression problem for non-zero flow magnitudes. Models are trained and evaluated using a spatial train–validation–test split based on Swiss cantons to assess generalization to unseen regions. Results show that machine learning methods substantially outperform traditional models, with CatBoost achieving the best classification performance and the neural network yielding the highest regression accuracy.**

I. Introduction

Human mobility plays a central role in shaping urban systems and has significant impacts on economic activity, environmental outcomes, transport demand, and spatial planning. Understanding and predicting commuting patterns is therefore essential for public policy, infrastructure development, and sustainable urban growth. As a result, human mobility has been widely studied, and various mathematical models have been proposed to describe and predict mobility flows.

Traditional approaches such as the gravity and radiation models [1], [2] explain commuting flows primarily through distance, population size, and spatial opportunity distributions. While these models are valued for their simplicity and interpretability, they rely on a limited number of parameters and fail to capture several socio-economic and urban factors influencing mobility. In recent years, machine learning methods have been increasingly applied to mobility modeling due to their ability to learn non-linear relationships from larger and heterogeneous datasets.

In this project, we study commuter flows between pairs of Swiss municipalities. The goal of the machine learning approach was to predict the spatial distribution of commuting flows using diverse urban, demographic, and socio-economic indicators. We implement and compare two traditional mobility models (gravity and radiation) with three machine learning approaches (XGBoost, CatBoost and a Fully Connected Neural Network model) and assess their relative strengths and limitations.

II. Traditional Methods

A. Gravitation method

The gavitational model is inspired by Newton's law of universal gravitation: it assumes that mobility decreases monotonically with distance and increases with population size. For a pair of municipalities (s, t) the flow of commuters from source municipality s to target municipality t is defined in the following way:

$$T_{st} = C\frac{m_s^\alpha n_t^\beta}{r_{st}^\gamma}$$

where m_s and n_t are the populations, while r_{st} is the distance between the two municipalities. Parameters C, γ, α and β can be found by applying ordinary Least Squares.

B. Radiation method

The radiation model relies on population distributions rather than distance alone. Mobility is constrained by the population of surrounding municipalities, which represents competing opportunities closer to the source. For any pair of municipalities (s, t), the commuter flow is defined as

$$T_{st} = T_s\frac{m_s n_t}{(m_s + p_{st})(m_s + n_t + p_{st})}$$

where T_{st} denotes the commuting flow, m_s and n_t are the respective populations, and p_{st} is the total population within a radius r_{st} centered at s. The model has a single parameter, T_s, representing the total outgoing commuters from s, which is estimated as $T_s = \alpha m_s$ using Ordinary Least Squares.

III. Data Preparation for Machine Learning

A. Prepare prediction data

1) Presence of only commuting flows: The commuting dataset contains observations only for municipality pairs with existing commuting flows. For machine-learning modeling, we completed the dataset by adding all remaining ordered municipality pairs and assigning them a commuting flow of zero.

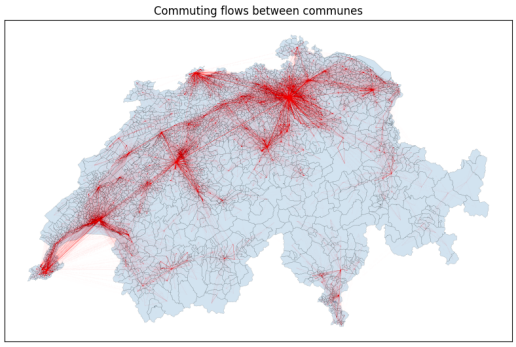

Figure 1. Commuting flow in Switzerland according to the Dataset [**?**]

2) Privacy constraints: Due to privacy constraints, commuting flows involving fewer than five individuals were aggregated into anonymous municipality codes labeled 7777. Since we cannot determine which municipalities are included in these aggregated codes and since they represent negligible flows, we chose to remove all observations containing the code 7777.

3) Definitions 'workplace' and 'residence place': The original dataset distinguishes two categories: workplace-based ("W") and residence-based ("R") commuting flows, differing only in the aggregation of low-flow observations. Since all 7777 entries were removed, the two datasets became identical. To avoid redundancy, we kept only the residence-based ("R") dataset for further analysis.

B. Feature preparation

For each ordered municipality pair $\langle m_i, m_j \rangle$, the feature vector is defined as $S_{ij} = \{d_{ij}, yr, pop_i, pop_j, U_i, U_j\}$. Where d_{ij} is the distance between the municipalities, yr the year, pop the population and U a set of urban indicators. The urban factors in our case are splitted between data we only found for the cantons, and the ones for the municipalities:

- **Canton**: GDP, Unemployment
- **Municipalities**: Traffic accidents, Age, Area, Population, Gender, Nationality.

The data for all features were collected from various Swiss data sources, each with different formats and preprocessing requirements.

1) Data of 2014: Most urban data for 2014 were not readily available online and had to be obtained from government agencies in highly heterogeneous formats. Harmonizing and preprocessing these datasets would have required substantial additional effort and computational resources. Also, the study focuses primarily on spatial rather than temporal analysis. For these reasons, the 2014 data were excluded in favor of prioritizing model development.

2) NAN Value treatment: Most NaN values resulted from differing data structures: some datasets identified municipalities by BFS numbers, others by name. Inconsistencies in name spelling (e.g., extra spaces) caused mismatches, particularly in the traffic accident dataset. To address this, missing values in the traffic accident dataset were imputed using a linear regression based on population. Remaining unmatched municipalities (30 in total) were dropped, as they represented a negligible fraction of the data.

3) Feature engineering: To capture non-linear and domain-specific commuting patterns, several engineered features were added. Distance decay was modeled using inter-communal distance and its squared term, while economic differences were represented by GDP differences between origin and destination. Population effects followed a gravity-model approach, using multiplicative interactions between origin and destination populations, with age-specific population shares included to reflect demographic influences. Additionally, a gravity-inspired index combining population and distance was constructed. Note that the neural network approach does not use these additional features.

Exploratory analysis of the engineered features revealed strongly skewed and non-linear distributions, motivating the application of logarithmic transformations to several features in order to stabilize variance and improve interpretability. This also contributed to the decision not to train the neural

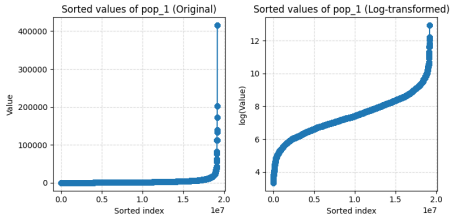

Figure 2. Caption

networks on the enhanced features, as it should be able to learn the relations between features itself. A similar transformation was applied to the target variable for the regression models. The response variable had a highly skewed distributon. Therefore, the regression models were trained on the log-transformed target values. For evaluation and interpretation, predictions were transformed back to the original scale.

IV. Model Preparation

A. General Pipeline

In order to properly handle the class imbalance and also to focus on the high variety of positive flows we decomposed the prediction task into two stages:

1) **Binary classification**: predicting whether a commuting flow exists between two municipalities.
2) **Regression**: estimating the magnitude of the flow conditional on its existence.

B. Train-validation-test split

The dataset was split using a custom spatial procedure instead of the standard *train_test_split*, to account for the strong spatial structure of commuting flows. Data were split by canton, training the models on all but one cantons and evaluating performance on an entirely unseen canton.

Luzern was chosen as the test set and Aargau as the validation set based on canton-level statistics. As the dataset is extremely sparse, with 98.69% of municipality pairs having zero flow, to choose these cantons we prioritized similarity to the global class imbalance as well as flow densities and socio-economic characteristics close to national averages. Luzern and Aargau have respectively 1.4% and 1.3% positive flows and represent 7.3% and 18.2% of all flows.

To handle class imbalance the splitting function is designed to drop a chosen portion of 0 flows. For CatBoost and XGBoost, zero-flow observations were retained in all splits, as these models handle class imbalance through class weighting. For the FCNN classifier, 90% of the zero-flow pairs were discarded. !!! justification

C. Grid-search

We performed hyperparameter tuning using a grid seach using three-fold cross-validation on the training data for Catboosting and XGBoost and a two-fold cross-validation for the neural network model. Because this procedure is computationally expensive, the search space was restricted to a limited set of the hyperparameters.
For XGBoost, as this model was faster than the two others, we were able to implement this grid search using a two-phase strategy: an initial randomized search was conducted over a coarse parameter space, followed by a finer grid search around best-peforming configurations.

Furthermore, since the default decision threshold for all three !!! models is 0.5 and therefore suboptimal in imbalanced settings, the classification threshold was optimized on the validation set to maximize the F1-score.

V. Models used

For the machine learning models used in this study, we selected algorithms that have been reported to perform well on similar mobility and flow prediction tasks in Brazil [3]. Furthermore, we included a neural network approach using a simple Fully Connected architecture with ReLU activations.

A. XGBoost

XGBoost is a tree-based gradient boosting framework known for strong performance on large, heterogeneous datasets and for capturing complex non-linear relationships [4].It has been shown to outperform many classifiers in commuter flow prediction tasks [3]. Since XGBoost does not natively handle categorical features, municipality codes were dropped and canton codes were one-hot encoded.

1) Classification: We trained an XGBoost classifier with a logistic objective on the train dataset. Class imbalance was handled using the `scale_pos_weight` parameter, set proportionally to the ratio of negative to positive samples. Model performance was assessed using the ROC-AUC metric, which is robust to class imbalance.

2) Regression: We train an XGBoost regressor to minimize the squared error loss, making it suitable for continuous target variables such as commuting flow volumes.

B. CatBoost

CatBoost, like XGBoost, is a gradient boosting decision tree algorithm. A key distinction is CatBoost's native support for categorical features (years and cantons in our study).

1) Classification: CatBoost was trained with the defined categorical features (cantons, years). Due to the class imbalance, the models performanced was assessed with F1-Score. The loss function used was Logloss. Class imbalance was further addressed by assigning higher training weights to municipalities with non-zero commuting flows via the class_weights parameter. A grid search over class weights and learning rate was conducted, but due to computational constraints, only a single grid search was performed without further refinement.

2) Regression: The grid search for the Regression was way faster, which is why we could do it over more parameters (depth, learning_rate, l2_leaf_reg).

C. Fully Connected Neural Network

For the neural networks, we used none of the categorical features. Note that we also logscaled all features, not just the heavily skewed ones. This was to keep all of them in roughly the same magnitude; Linear scaling would have been an option as well.

1) Classification: Like CatBoost, the classification network was trained with binary cross-entropy loss. To improve performance on the positive class and reduce training time, 90% of zero-flow entries were discarded. Hyperparameter search was performed on a small subset,

tuning learning rate, weight decay, and model size, while keeping the batch size fixed at 2048. In the final training, early stopping based on ROC-AUC was used to prevent overfitting, and the classification threshold was optimized for F1, chosen over accuracy due to strong class imbalance.

2) Regression: For the regression, we used MAE loss. As we ignored zero-flow entries for the regression, the dataset was much smaller, so we used all entries for the hyperparameter search. We evaluated combinations of learning rate and weight decay. Batch size was again fixed, now at 128. Training our chosen model, we used a patience mechanism again, on the validation loss.

VI. Results on the Test Dataset

A. Traditional methods

Traditional methods were trained exclusively on non-zero flows as the logarithm cannot take 0 as an argument. We trained on every canton except Luzern and tested on it, the results are illustrated in the table below.

Table I
F1 Scores for rational and gravitational method

Method	R^2	Correlation
Radiation method	0.128	0.39
Gravitation method	0.165	0.74

The traditional methods performed poorly compared to our machine learning models, consistent with [3]. While their R^2 does not exceed 0.2, the ML models reach up to 0.84, demonstrating substantially superior predictive performance.

B. Different ML models

1) Classification: The results for the different methods on classification are displayed in the table below. CatBoost performs slightly better than XGBoost, likely because it can incorporate categorical features such as cantons. Spatially, the model predicts local commuting well but struggles with long-distance flows. Including data on main train lines [5] and highways could likely improve long-distance predictions.

Table II
Classification: Hyperparameter settings and F1 scores. For all hyperparameters of XGBoost, see in the code

Method	Hyperparameter value	F1 score
XGBoost	Depth: 5 Learning rate: 0.2 ...	0.68
CatBoost	learning rate: 0.1 class weights: [1, 5] depth: 8	0.70
FCNN	# nodes in each layer: 64,48,48,32,1 batch size: 2048 learning rate: 1e-4 weight decay: 1e-5	0.67

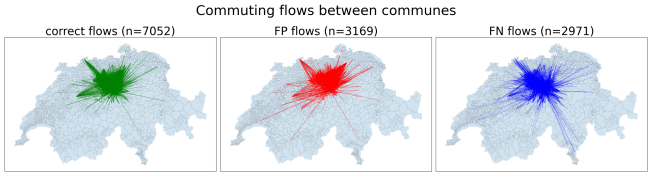

Figure 3. Different results for the classification algorithm of CatBoost

2) Regression: For the regression task, the CatBoost model performed the worst, whereas the CNN achieved the best results. This is likely because the CNN does not rely on manual feature engineering and can learn complex patterns directly from the data. The relatively poorer performance of CatBoost compared to XGBoost is also consistent with observations reported in our reference study [3].

Table III
Regression: Hyperparameter settings and R^2 scores. For all hyperparameters of XGBoost, see in the code

Method	Hyperparameter value	R^2 score
XGBoost	Depth: 0.44 Learning rate: 0.08 ...	0.78
CatBoost	learning rate: 0.1 l2_leaf_reg: 1 depth: 8	0.34
FCNN	# nodes in each layer: 48,48,32,32,1 batch size: 128 learning rate: 1e-4 weight decay: 1e-6	0.84

VII. Conclusion

Our results confirm the findings of [3] that machine learning models outperform both gravity and radiation models. We also demonstrated that fully connected neural networks (FCNNs) provide a viable alternative for predicting flows between municipalities, but still work less well for some of the tasks such as the classification.

For future research, model performance could be improved through a more extensive hyperparameter search and, for CNNs, more complex architectures. Incorporating additional data, such as train lines and highways, could further enhance predictions. Alternatively, these models could be adapted to temporal tasks, such as forecasting future commuting flows rather than predicting flows in unseen municipalities.

REFERENCES

[1] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, vol. 734, pp. 1–74, 2018, human mobility: Models and applications. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037015731830022X

[2] M. Stefanouli and S. Polyzos, "Gravity vs radiation model: two approaches on commuting in greece," *Transportation Research Procedia*, vol. 24, pp. 65–72, 2017, 3rd Conference on Sustainable Urban Mobility, 3rd CSUM 2016, 26 – 27 May 2016, Volos, Greece. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146517303502

[3] G. Spadon, A. C. P. L. F. de Carvalho, J. F. Rodrigues-Jr, and L. G. A. Alves, "Reconstructing commuters network using machine learning and urban indicators," *Scientific Reports*, vol. 9, no. 1, p. 11801, 2019.

[4] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[5] Swiss Federal Railways (SBB), "Sbb's route network," https://data.sbb.ch/explore/dataset/linie/information/, 2025, accessed: 2025-12-XX; Open Data from data.sbb.ch – SBB Open Data platform.

DIGITAL ETHICS CANVAS

CONTEXT

Predicting commuting movements in Switzerland

SOLUTION

A machine learning model predicting commuting flows between Swiss communes to support transport and infrastructure planning

BENEFITS

The project can have different benefits:
- Improve transport planning and infrastructure decisions by anticipating flows
- Allow interventions to reduce congestion, promote equitable mobility, and support sustainable transport.
- Insights for environmental impact (e.g., emissions reduction via planning)
- Research on population movements and regional development.

WELFARE

RISK

- Can the solution be used in harmful ways, in particular with regards to vulnerable populations?
- What kind of impacts can errors from the solution have?
- What type of protection does the solution have against attacks or misuse?

- Results might be used to justify infrastructure developpement in some regions and not others, for example well-connected or data-rich urban areas may benefit more than rural or peripheral regions.
- Bad prediction and construction of infrastructure might lead to not adapted infrastructure (overflow, underflow, increase in travel time), and inequality in those.
- Predictions based on historical data might reinforce existing mobility patterns instead of enabling change

MITIGATION

- Evaluate model based on region size and socioeconomic factors
- Highlight in the solution for what regions the model is working well and which not
- Use model outputs as decision support, not as sole justification for infrastructure development; combine results with expert judgment and local knowledge.

FAIRNESS

RISK

- How accessible is the solution?
- What kinds of biases may affect the results ?
- Can the outcomes of the solution be different for different users or groups?
- Could the solution contribute to discrimination against people or groups?

- Smaller municipalities with less data have less good predictions
- Model works better on closer distances => longer distances but popular commuting ways such as Ticino - Zürich might be overlooked, infrastructure underdevelopped therefore
- Flows from less connected regions are underrepresented => might especially be interessting to develop infrastructure there (eg. some urban regions...)

MITIGATION

- Apply different models for municipalities with less data
- Add data about main train lines already existing to take into account commuting movement over longer distances
- Add a feature of what reasons it might be interesting to promote commuting movement (high population, but bad connection, urban regions, ...)

AUTONOMY

RISK

- Can users understand how the solution works and what its limits are?
- Are users able to make choices (e.g. consent, settings) in their use of the solution and how?
- How does the solution affect user autonomy and agency?

- Used in applications etc, users would use it without understanding whats behind.
- Used by train companies / decision making entities, important to be aware of features used for the model

MITIGATION

- Clear structured data, clearly explained, model public

PRIVACY

RISK

- What data does the solution collect
- Is it collecting personal or sensitive data
- Who has access to the data?
- How is the data protected?
- Could the solution disclose / be used to disclose private information?

- Especially for data with low commuting movement, the numbers can be linked back to the person - violating their privacy

MITIGATION

Privacy issues were taken into account through the category '7777' in the dataset - data anonymized and finally dropped

SUSTAINABILITY

RISK

- What is the carbon footprint of the solution?
- What types of resources does it consume (e.g. water) -and produce (e.g. waste)?
- What type of human labor is involved?

- If new street / trainlines are created based on these solution, it does have a considerable environmental impact.
- If those streets / trains are afterwards not used accordingly,
it does have a negative environmental impact with no benefits, which wouldn't be wished for
- Using the models for justifying the creation of new roads would have a negative carbon impact.

MITIGATION

The ML work need to be completed with other studies, people working more in this domain and having more experience about how and where new trianlines / streets would need to be created.

Digital Ethics Canvas 2024 - C. Hardebolle, V. Macko, V. Ramachandran, Y. Do, N. Bartłomiejczyk, A. Holzer, P. Jermann