Non-linear associations between human mobility and sociodemographic and contextual factors

Milad Malekzadeh[1], Jed A. Long[2]

[1]Department of Geography and Environment, Western University, mmalekz4@uwo.ca
[2]Department of Geography and Environment, Western University, jed.long@uwo.ca

ABSTRACT

Daily human mobility patterns are influenced by different factors, such as contextual (e.g., underlying transportation network and land use) and sociodemographic factors. The relationship between these factors and human mobility can be complex and non-linear. However, researchers typically study these relationships by using linear models. We argue that linear models may lead to misinterpretation; hence, we investigate the complexity of these relationships using non-linear models. In this study, we explore 20 factors at aggregated level and their non-linear associations with a measure of mobility (Radius of Gyration) using a random forest machine learning model. We used an available mobility dataset in Ontario, Canada, from our Ontario COVID-19 Mobility Dashboard. Our results show complex, non-linear, and non-monotonic relationships between factors and human mobility. Variables relating to sociodemographic factors had the highest importance level. We compare with a linear regression model to assess our results and observe a clear contrast with the random forest model's results. Our results indicate that while linear models are commonly used in such studies, the interpretations might be erroneous. We suggest that the relationship between human mobility and such factors should be examined by models capable of capturing non-linearity.

1. Introduction:

Human mobility is influenced by different factors, such as sociodemographic and environmental factors (Chakrabarti et al., 2021; Firth et al., 2022; Luo et al., 2016; Pappalardo et al., 2015; Ruktanonchai et al., 2021). Sociodemographic factors such as age, income, and education (Schwanen et al., 2002; Stopher et al., 2003; Volosin et al., 2013) are commonly considered in the literature. There is also broad attention in the literature on how contextual factors which shape our living environment impact mobility patterns such as the underlying transportation network and facilities which influence travel time, travel mode, number of trips, and the time we allocate to our activities (Schwanen et al., 2002). Understanding the relationships between these factors and human mobility enables us to make more informed decisions in transportation planning, urban planning, public health, and environmental science (Hidayati et al., 2021; Lenormand et al., 2015).

Researchers have been trying to model and investigate the relationships between these factors and human mobility for decades. However, since linear regression models are easy to implement and interpret, they are commonly employed for studying relationships between human mobility patterns and sociodemographic and contextual factors. Non-linear machine learning methods, on the other hand, are capable of capturing if non-linearities are present in the relationships between mobility and sociodemographic and contextual factors. However, they are

 Volume DOI: 10.17605/OSF.IO/N6DU5

more complex and usually are considered black boxes that are difficult to interpret (Xin et al., 2022).

Over the last few years, studies have been beginning to use machine learning models, largely for the purpose of *predicting* human mobility patterns. For instance, researchers have been trying to use machine learning methods to model and predict human mobility using a variety of different approaches (Lee et al., 2010; Monreale et al., 2009; Song et al., 2017). However, there was less attention to interpreting the models and the relationships between mobility and sociodemographic and contextual factors.

In this study, we apply machine learning models (i.e., a random forest model) to study the non-linear relationships between mobility patterns and sociodemographic and contextual factors. We use in total 20 different explanatory variables in the model. To evaluate and interpret the random forest model we use accumulated local effects plots and permutation feature importance. We compare the random forest model with a classical linear regression.

2. Data:

Data on human mobility was sourced from the Ontario COVID-19 Mobility Dashboard (Geospatial Analysis Lab, 2021; Long et al., 2021; Long & Ren, 2022). We used the human mobility data from 02 February 2020 to 08 February 2020 as an example.

We derived sociodemographic data from the Statistics Canada data source (Statistics Canada, 2017) through the Esri Enrich layers tool (ESRI, 2022). We employed ratio of children at home, the ratio of worker population, total money spent on transportation per person, total money spent on public transportation per person, total money spent on private transportation per person, total money spent on gas per person, total money spent on recreational activities per person, age, the ratio of detached house owners, the ratio of post-secondary educated, the median income of households, and the ratio of visible minorities in our study.

We chose to use eight factors that relate to the context of a region, which we expected to be related to observed mobility patterns. The eight factors we chose were: population density from Statistics Canada (Statistics Canada, 2017), the Walk score, public transit score, and bike score, extracted from (Walk Score, 2014); intersection density and road density as a proxy of the transportation network, and points of interest density from OpenStreetMap data (OpenStreetMap contributors, 2017); built environment ratio to compare the built environment areas to other land uses (Agriculture and Agri-Food Canada, 2022) which was originally derived from Landsat5-TM and 7-ETM+ multi-spectral imagery.

3. Methods:

We used the radius of gyration (ROG) which is a measure of the range of mobility, previously used in numerous studies (González et al., 2008).

$$ROG = \sqrt{\frac{\sum_i^n (d_{i,c})^2}{n}} \quad (1)$$

Where $d_{i,c}$ is the distance from point i to central point c (here, home neighbourhood location; see Long & Ren 2022), and n is the number of stops.

To avoid issues associated with multicollinearity in the models, we remove highly correlated factors. We computed the Pearson correlation of all factors, and based on Cohen's

 Volume DOI: 10.17605/OSF.IO/N6DU5

relationship strength (Cohen, 2013), we removed the minimum number of factors with which we could avoid having a correlation (Pearson-R) over 0.7. From the initial 20 factors, we eliminated 6 factors: total money spent on transportation per person, total money spent on private transportation per person, total money spent on recreational activities per person, public transit score, bike score, and road density. To employ these factors in machine learning models, we standardized the values of the factors to avoid the dominancy of any factors that have larger scales.

After preliminary testing between a support vector regression (SVR), a random forest (RF), and an artificial neural network (ANN), we determined that the RF model was the best option for studying non-linear relationships between human mobility and sociodemographic and contextual factors. We implemented an RF model using the Scikit-learn package in Python (Pedregosa et al., 2011). We chose the RF model with 39 trees since it had the lowest root mean square error. We set the minimum number of samples required to be at a leaf node to 4 to avoid overfitting.

To assess the performance of the RF model, we divided our dataset into training data (70%) and test data (30%). We computed the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination/efficiency (R2) of the logarithm of estimated ROG values for both the training data set and validation data set. A classical linear regression model was also fitted to enable us to compare the results of the RF model.

To measure the influence of each factor in the models, we implemented a model-agnostic evaluation method on both models - the permutation feature importance (PFI) algorithm (Altmann et al., 2010). PFI is a measure of decrease in accuracy (DA) when the feature value is changed. To assess the difference between the magnitude of the influence of each factor in the two models, we ranked variables in each. To visualize the relationship between the factors and human mobility, we implemented accumulated local effects (ALE) plots (Apley & Zhu, 2020). While partial dependence plots (PDP) (Friedman, 2001), are the most popular visualization method to present the individual effects of different variables, they might produce erroneous results especially when the variables are highly correlated. On the other hand, ALE plots are unbiased alternatives to PDPs. ALE plots average the changes in the dependent variable where changes are computed only for a specific window around a value. This eliminates the bias produced by unrealistic instances in PDP calculation. ALE plots demonstrate the marginal effect of each variable on the dependent variable. Therefore, ALE plots will enable us to better understand the potential non-linear pattern of relationships between the different factors and human mobility, as measured by ROG.

4. Results:

The RF model had a much lower error level in the training data set than the linear regression model based on the MAE, RMSE, and R2 values (Table 1). The RF model's error results in the test dataset are again lower than the linear regression, however, in the test data, this difference is much smaller (Table 1).

Table 1 - Error analysis results comparing a random forest (RF) and Linear Regression model predicting human mobility (Radius of Gyration) from a set of 20 predictor variables.

	Training			**Test**		
	MAE	RMSE	R2	MAE	RMSE	R2
Linear Regression	0.201	0.257	0.421	0.198	0.253	0.356
RF	0.109	0.148	0.816	0.198	0.251	0.370

4 Non-linear associations between human mobility and sociodemographic and contextual factors

In the RF model, the highest importance score was associated with the variable median income (Table 1), more than twice higher than the second variable. The ratio of children at home was ranked as the second most important variables in the RF model, but the decrease in accuracy score, was less than ½ of the median income variable (Figure 1). Population density and POI density were the highest ranking contextual variables (DA = 0.012 and 0.011, respectively). These results suggest that in the RF model, three socio-deomographic variables were the three most important ones (median income, the ratio of children at home, and the ratio of visible minority). In the linear regression model, the ratio of children at home was the most important factor (Figure 1). In the linear regression model, the built environment ratio variable was the highest ranking contextual variable (DA = 0.007). Two variables that differed greatly in importance between the two models were the age and population density factors. The age factor had low importance in the RF model (DA = 0.005) and high importance in the linear regression model (DA = 0.005). Population density factor had high importance in the RF model (DA = 0.012) and low importance in the linear regression model (DA = 0.001).

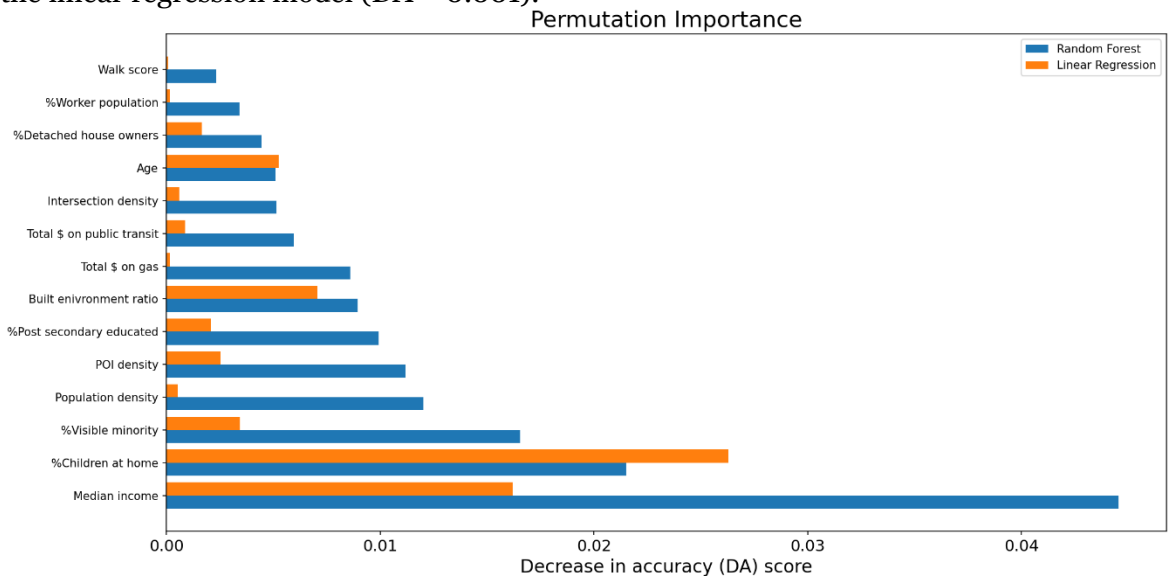

Figure 1 – Permutation importance plot indicating the level of influence of each factor on human mobility in the random forest model (blue) and the linear regression model (orange) – demonstrating the amount of decrease in accuracy (DA) score (increase in RMSE).

To assess the structure of relationships between the factors in the models and human mobility, we used ALE plots (Figure 2). For most factors, the relationship is complex, non-monotonic, and non-linear. For instance, we observe a non-montonic, non-linear association between both ratio of detached house owners and POI density and human mobility; whereas the linear regression model identifies a positive linear trend. The POI density plot is positively skewed, and in the lower end of the distribution where most of the data lies, we observe a negative association in the RF model, whereas, in the same range of values, the linear regression model demonstrates a positive association (Figure 2).

Volume DOI: 10.17605/OSF.IO/N6DU5

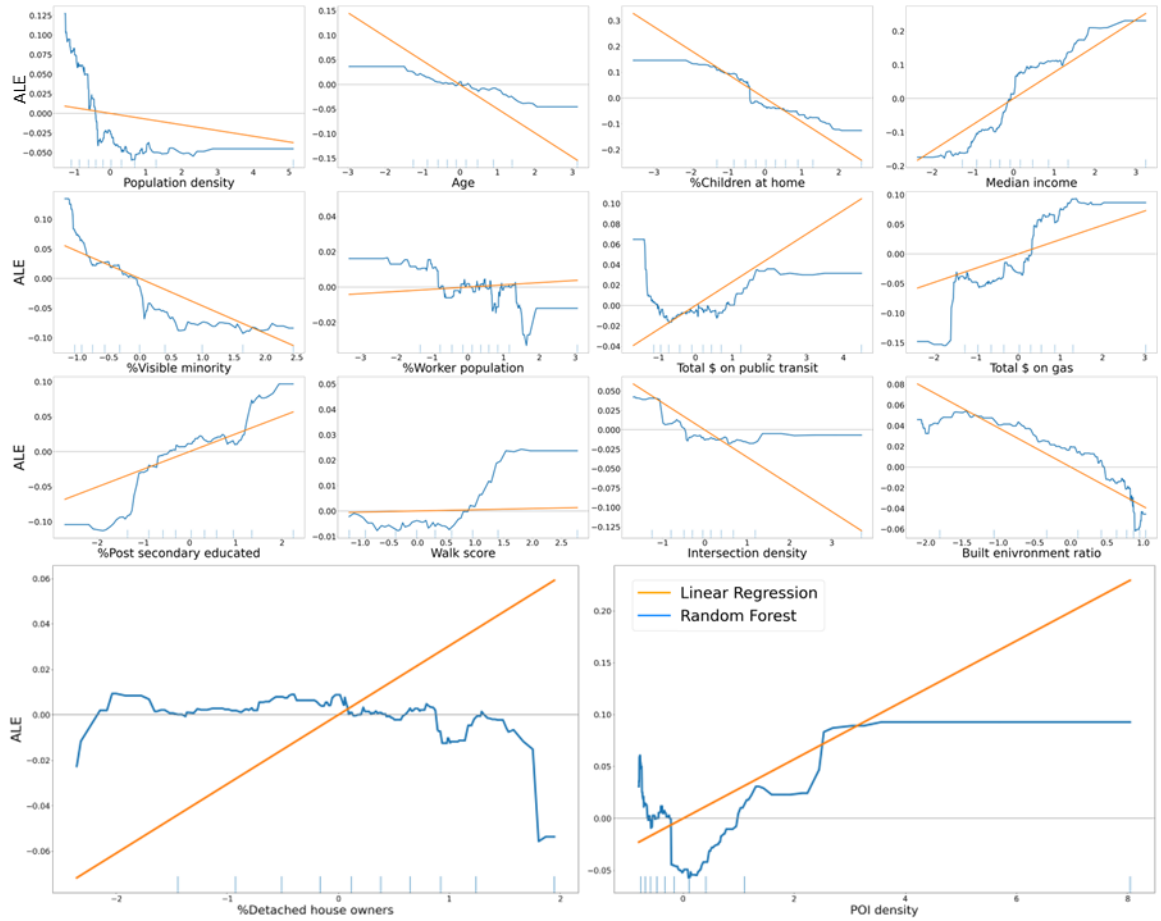

Figure 2 – Accumulated Local Effects (ALE) plots of variables – ALE against standardized factors' values. The blue and yellow lines are the RF and the linear regression model, respectively. Vertical short lines at the bottom indicate each decile of the underlying data distribution.

5. Discussion & Conclusion:

Understanding the association between sociodemographic and contextual factors and human mobility is of importance for urban planners, transportation planners, and decision-makers. Focusing on their non-linearity, we demonstrated that previously used linear regression models are not proper models especially when non-linear associations are present and non-judicious use of linear models might result in erroneous interpretations and misjudgment of the associations. Using the ALE plots, we clearly demonstrated that the linear regression model did not capture non-monotonic and/or non-linear associations. Specifically, this was most apparent in the variables: population density, the ratio of worker population, total money spent on public transit, the ratio of detached house owners, and POI density. We also observed contrasting results in PFI analysis as each factor's degree of importance and their ranks changed in the models but overall, median income and % children at home with the two most important variables in both models.

Although we tried to examine a large number (n=20) of easily computed factors, we did not comprehensively cover all possible factors related to human mobility We used the POI data set from OpenStreetMaps which does not include a measure of the attraction level of POIs; nor can be considered a fully comprehensive data set on POIs. Similarly, we used aggregated-level data, and therefore we did not have individual level information which could significantly impact

human mobility (D. Y. Kim & Song, 2018; S. Y. Kim et al., 2016). This means that although all of what we discussed can be true at an aggregated level, our inferences should not be generalized to individual behaviors.

In conclusion, we highlighted that non-linear associations are present when studying human mobility patterns. We also found that, at the aggregate level, sociodemographic factors seem to be more influential relative to contextual factors. With reference to this example, emphasizing the importance of non-linearity in associations, we provided a workflow of the model selection procedure for studying human mobility and the factors that influence it.

Acknowledgments:

The project was supported by a Western Catalyst Grant, and from the Natural Sciences and Engineering Research Council of Canada.

References:

Agriculture and Agri-Food Canada. (2022). *Land Cover for Agricultural Regions of Canada*. Canada Open Government Website. https://open.canada.ca/data/en/dataset/16d2f828-96bb-468d-9b7d-1307c81e17b8

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(4), 1059–1086.

Chakrabarti, S., Hamlet, L. C., Kaminsky, J., & Subramanian, S. v. (2021). Association of human mobility restrictions and race/ethnicity–based, sex-based, and income-based factors with inequities in well-being during the COVID-19 pandemic in the United States. *JAMA Network Open*, *4*(4), e217373–e217373.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

ESRI. (2022). *Enrich (Analysis) - ArcGIS Pro 3.0*. ESRI Company. https://pro.arcgis.com/en/pro-app/latest/tool-reference/analysis/enrich.htm

Firth, C. L., Kestens, Y., Winters, M., Stanley, K., Bell, S., Thierry, B., Phillips, K., Poirier-Stephens, Z., & Fuller, D. (2022). Using combined Global Position System and accelerometer data points to examine how built environments and gentrification are associated with physical activity in four Canadian cities. *International Journal of Behavioral Nutrition and Physical Activity*, *19*(1), 1–12.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

Geospatial Analysis Lab. (2021). *Ontario COVID-19 Mobility Dashboard* . Western University. https://geospatial.uwo.ca/mobility.html

González, M. C., Hidalgo, C. A., & Barabási, A. L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782. https://doi.org/10.1038/nature06958

Hidayati, I., Tan, W., & Yamu, C. (2021). Conceptualizing mobility inequality: Mobility and accessibility for the marginalized. *Journal of Planning Literature*, *36*(4), 492–507.

Kim, D. Y., & Song, H. Y. (2018). Method of predicting human mobility patterns using deep learning. *Neurocomputing*, *280*, 56–64.

Kim, S. Y., Koo, H. J., & Song, H. Y. (2016). A study on influence of human personality to location selection. *Journal of Ambient Intelligence and Humanized Computing*, *7*(2), 267–285.

Lee, J.-G., Han, J., Li, X., & Cheng, H. (2010). Mining discriminative patterns for classifying trajectories on road networks. *IEEE Transactions on Knowledge and Data Engineering*, *23*(5), 713–726.

Lenormand, M., Louail, T., Cantú-Ros, O. G., Picornell, M., Herranz, R., Arias, J. M., Barthelemy, M., Miguel, M. S., & Ramasco, J. J. (2015). Influence of sociodemographic characteristics on human mobility. *Scientific Reports*, *5*(1), 1–15.

Long, J. A., Malekzadeh, M., Klar, B., & Martin, G. (2021). *Do regionally targeted lockdowns alter movement to non-lockdown regions? Evidence from Ontario, Canada*. 102668. https://doi.org/10.1016/j.healthplace.2021.102668

Long, J. A., & Ren, C. (2022). Associations between mobility and socio-economic indicators vary across the timeline of the Covid-19 pandemic. *Computers, Environment and Urban Systems*, *91*, 101710.

Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography*, *70*, 11–25.

Monreale, A., Pinelli, F., Trasarti, R., & Giannotti, F. (2009). Wherenext: a location predictor on trajectory pattern mining. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 637–646.

OpenStreetMap contributors. (2017). *Planet dump retrieved from https://planet.osm.org* .

Pappalardo, L., Pedreschi, D., Smoreda, Z., & Giannotti, F. (2015). Using big data to study the link between human mobility and socio-economic development. *2015 IEEE International Conference on Big Data (Big Data)*, 871–878.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Ruktanonchai, C. W., Lai, S., Utazi, C. E., Cunningham, A. D., Koper, P., Rogers, G. E., Ruktanonchai, N. W., Sadilek, A., Woods, D., & Tatem, A. J. (2021). Practical geospatial and sociodemographic predictors of human mobility. *Scientific Reports*, *11*(1), 1–14.

Schwanen, T., Dijst, M., & Dieleman, F. M. (2002). A microlevel analysis of residential context and travel time. *Environment and Planning A*, *34*(8), 1487–1507.

Song, X., Shibasaki, R., Yuan, N. J., Xie, X., Li, T., & Adachi, R. (2017). DeepMob: learning deep knowledge of human emergency behavior and mobility from big and heterogeneous data. *ACM Transactions on Information Systems (TOIS)*, *35*(4), 1–19.

Statistics Canada. (2017). *Ontario [Province] and Canada [Country] (table). Census Profile 2016 Census*. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/Page.cfm?Lang=E&Geo1=PR&Code1=35&Geo2=&Code2=&SearchText=Ontario&SearchType=Begins&SearchPR=01&B1=All&GeoLevel=PR&GeoCode=35&type=0

Stopher, P. R., Greaves, S., & Bullock, P. (2003). Simulating household travel survey data: application to two urban areas. *82nd Annual Meeting of the Transportation Research Board, Washington, DC*.

Volosin, S. E., Paul, S., Christian, K. P., Konduri, K. C., & Pendyala, R. M. (2013). Exploring the dynamics in travel time frontiers. *Transportation Research Record*, *2382*(1), 20–27.

Walk Score. (2014). Walk score methodology. *Accessed April*, *24*.

Xin, Y., Tagasovska, N., Perez-Cruz, F., & Raubal, M. (2022). Vision paper: causal inference for interpretable and robust machine learning in mobility analysis. *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–4.