GENERATIVE DESIGN FOR PRECISION GEO-INTERVENTIONS

R. Wen [1], S. Li [1]

[1] Department of Civil Engineering, Faculty of Engineering and Architectural Science, Toronto Metropolitan University, 350 Victoria Street, Toronto, Canada - (rwen, snli)@ryerson.ca

KEY WORDS: Interventions, GIS, Decision Support, Generative Design, Machine Learning, AutoML, Geodesign, Optimization

ABSTRACT:

The purpose of this research is to develop an approach for a Spatial Decision Support System (SDSS) that integrates Geographic Information Systems (GIS), Automated Machine Learning (AutoML), and Hyperparameter Optimization (HPO) to generate precision geo-interventions based on standardized geospatial data and user design constraints. The geo-intervention generation approach involves three steps: (1) Geo-binning, (2) AutoML, and (3) Prediction Optimization. Geo-binning is used to standardize geospatial data into regularized grids as inputs into AutoML models. Prediction optimization generates geo-interventions by applying user-design constraints and optimizing AutoML model output to find optimized input variables that form precise geo-interventions. An experiment in reducing road traffic collisions using infrastructural changes in Toronto, Ontario, Canada was done to evaluate the geo-intervention generation approach. The results of the experiment found that changing the number of schools, red light cameras, and transit shelters in high traffic areas could potentially halve the total number of traffic collisions according to a 80 by 80 geo-binned grid Auto-Sklearn model with a Mean Absolute Error (MAE) of 117.68. It was also found that user design constraints heavily affected the prediction optimization step as when the areas were altered to an alternative grid of cells with scarce infrastructure, the number of predicted collisions rose by 6127 collisions. Thus, limitations of this study included subjectivity in user design constraints, scalability, and interactivity. Future work involves improving modelling/optimization efficiency and developing an interactive interface for exploring generated precision geo-interventions.

1. INTRODUCTION

Geo-interventions, actions implemented in geographic space that alter specific outcomes (e.g., safe road design for reducing traffic collisions and hotspot policing for reducing crime), are an effective solution to reducing a large portion of injuries from traffic collisions and violent crimes, which typically occur in large urban settings. Common approaches (e.g., cluster mapping, cellular automata, and multiple criteria decision analysis) to modelling and analysing geo-interventions have focused on identifying target areas/risk factors and simulating scenarios/impacts/theories to support decision-making (Malczewski and Rinner, 2015). However, these approaches usually model/analyse a range of existing/pre-defined geo-interventions, as opposed to generating or exploring potentially new geo-interventions based on data and user design constraints (Mehaffy, 2008).

In addition, common approaches to modelling/analysing geo-interventions rely heavily on domain expertise (e.g., selecting/interpreting models/variables, data processing, and model assumptions) and evaluate only a small number of alternative geo-interventions (typically ranging from perhaps 3 to 12 alternatives). With recent advancements in large-scale computing and data availability in urban settings, there is huge potential to explore hundreds to thousands of alternative urban geo-interventions (Li et al., 2016). This reduces the heavy reliance on domain expertise by exploring a larger space of alternatives with computing power and big urban data, which leads to substantially more comprehensive experiments and impact evaluations.

AutoML has had great success in using large-scale computing and big data to automatically pre-process data, select important variables, and discover/compare accurate models across large search spaces (He et al., 2021). Hyperparameter Optimization (HPO) has also been effective at improving model performance in AutoML approaches through the optimization of model parameters given constraints such as time, parameter ranges, or desired performance criteria (Feurer and Hutter, 2019).

This research proposes an approach for a SDSS that integrates AutoML and HPO with GIS to leverage modern advancements in computing power and big data availability. Spatial binning, a GIS technique, is first used to standardize and aggregate the geospatial data into polygonal bins (e.g., cells and hex-grids). AutoML is then used to automatically pre-process and generate geo-intervention models based on existing geospatial data. HPO is finally used to optimize the most performant models from the AutoML process under user design constraints (e.g., applicable intervention areas, budget/resource constraints, and desired impact). In the HPO process, model inputs represent the potential geo-interventions (e.g., road width and number of traffic speed cameras/schools/police stations), while the outputs represent the predicted impacts from the geo-interventions (e.g., change in traffic collisions/stabbings/gun violence). By optimizing the inputs to the AutoML models, the HPO process explores and generates hundreds to thousands of possible geo-interventions based on the model inputs and outputs automatically. These potential geo-interventions are precise – locatable to each grid cell based on the spatial binning resolution (e.g., grid size), and quantitatively measured as a change in specific model inputs for each cell, which can be visualized as GIS map layers.

2. METHODS

The method for generating geo-interventions consist of three steps:

(1) Geospatial Binning (Geo-binning)
(2) Automated Machine Learning (AutoML)
(3) Prediction Optimization

The initial input consists of any geospatial data in a standard format read by the *geopandas*[1] Python library. This data is standardized into regularized grid cells by binning, and then aggregating, the geospatial data and their associated variables by each grid cell. This regularized grid is then used as input for the AutoML step, which involves a target variable y (the variable to be predicted) and inputs $x_1 \ldots x_n$. After building an AutoML model with adequate performance, the predictions from the model are optimized based on user design constraints relevant to a decision-making problem. For example, if there is suspicion that red light cameras help reduce the number of traffic collisions, the AutoML model would produce predictions on the number of traffic collisions in each cell, while the user may enforce that only areas without red light cameras and high traffic can be changed. Thus, the prediction optimization step would try to minimize the number of collisions while attempting to determine the number of red light cameras needed for each cell to help reduce traffic collisions in the desired areas. The overall workflow of the methods in this paper are illustrated in Figure 1.

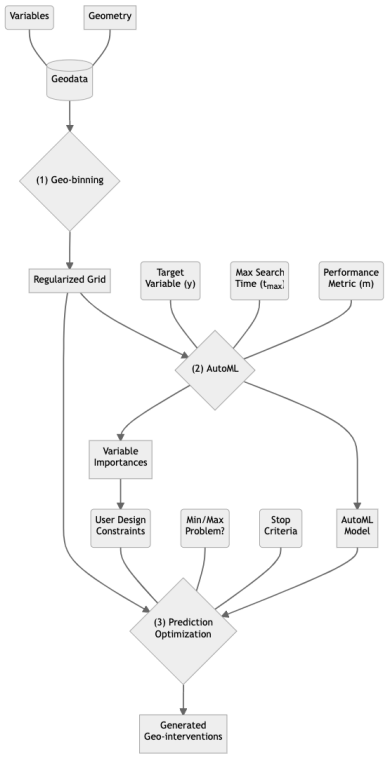

Figure 1. Methods for Generating Geo-interventions.

[1] https://geopandas.org

2.1 Geospatial Binning (Geo-binning)

A regularized grid of cells is used to bin and aggregate input geospatial data before the AutoML step. This standardizes the input data for the AutoML models into each cell of the grid by aggregating input geospatial geometries and associated variables contained inside each cell. When aggregating geometries and variable values, numeric statistics are calculated and stored inside each cell. These statistics include the: sum, mean, median, min, max, variance, skew, standard deviation, standard error of the mean, and mean absolute deviation. Geometric data are aggregated for each cell if they intersect from applying spatial joins. The aggregation behaviour applied is based on the three basic geometry types seen in Table 1.

Geometry Type	Aggregation Behaviour
Point	Count points inside cell
Line	Count line objects inside cell; Calculate statistics for line lengths and sinuosity in cell
Polygon	Count polygon objects inside cell; Calculate statistics for polygon areas, lengths, and widths in cell

Table 1. Geo-binning behaviour for geometry types.

Variables inside each geospatial dataset are assumed to be numeric or textual. Given this assumption, the aggregation behaviour applied to variable values is based on the two generic data types seen in Table 2. When the geometries and variables are aggregated into a regularized grid of cells, each cell can be seen as a row of data, while the aggregated statistics and counts are columns in a standard tabular format. This allows using this data for training and testing any model that accepts tabular formats.

Data Type	Aggregation Behaviour
Numeric	Calculate statistics for variable values in cell
Textual	Count unique variable values in cell

Table 2. Geo-binning behaviour for data types.

2.2 Automated Machine Learning (AutoML)

After geo-binning the geospatial data into a regularized grid of cells where each row represents a cell and each column represents aggregated statistics for geometries or variables inside each cell, the geo-binned data are used as training data to create AutoML models. A target variable y must be specified as the variable to be predicted while all other variables $x_1 \ldots x_n$ are considered independent variables used to predict y. The modelling problem then becomes $y_p = f(x_1 \ldots x_n)$, where y_p is the predicted target variable from the model. An appropriate metric $m(y, y_p)$ representing the performance of the model is then used to guide the AutoML process to find the most accurate model provided with a time constraint $t \leq t_{max}$. Thus, the user is required to only define the target variable y, the metric m (can be selected from pre-determined standard metrics), and the maximum running time t_{max} allowed to search for an AutoML model.

Two approaches were used in this paper to produce AutoML models: (1) A genetic algorithm-based approach called Tree-Based Pipeline Optimization Tool (TPOT) using the Python package *tpot* and (2) A Bayesian optimization based approach called Auto-Sklearn using the Python package *autosklearn*.

Commented [A1]: Each row represents a collection of cells rather than a cell?

Commented [RW2R1]: Each row represents a cell, each column is a variable, one cell has multiple variables

TPOT evolves tree-based machine learning pipelines with genetic programming to automate variable selection, transformation, construction, and model selection and parameter optimization (Olson et al., 2016). Auto-Sklearn initiates a suggested machine learning pipeline using meta-learning, then optimizes this pipeline of data/feature pre-processors and a known set of models followed by ensembling these pipelines to produce AutoML models (Feurer et al., 2015). For both approaches, regression and classification models are both available, and no additional manual processing of the geo-binned grid is required as input to these AutoML models.

In addition to creating AutoML models, permutation importance is calculated to interpret the effects of variables in AutoML models. Permutation importance measures variables affect the model performance if variables are randomly shuffled, which indicates how important variables are to any generic model (Altmann et al., 2010). The mean permutation importance over randomly shuffled runs is used to guide the user design constraints for the prediction optimization step, allowing the user to select particular variables to generate geo-interventions for more important variables that indicate significant areas for intervention and geospatial objects that have larger effects on the target variable *y*.

2.3 Prediction Optimization

After creating an AutoML model, geo-interventions are generated by optimizing the predictions and inputs for this model. User design constraints related to the filtering of data for optimization are used to guide the optimization process, provide better initial starting points, and reduce the search space for efficiency (Feurer and Hutter, 2019). The user may limit the optimization process by only particular variables and to only specified cells in the geo-binned grid. These design constraints can be guided by the variable importance from the AutoML models. For example, the optimization process can be limited to the top three most important variables from the permutation importance computations using the AutoML models. To further this example, grid cells can be limited to priority areas in which one or more of the important variable values are much higher or lower than the average. After filtering the input variables and grid cells for optimization, Bayesian Optimization is used to find the optimal input values for each grid cell to maximize or minimize an aggregate metric based on the predicted target *y* values (Wu et al., 2019). Since the major consideration and constraint often relates to the time allowed for the search, Bayesian optimization offers an appropriate approach for optimization by using Gaussian processes and previous samples to lower the number of iterations and relatively reduce the time needed to find more optimal values (Snoek, Larochelle, and Adams, 2012). When the optimal aggregated variable inputs and associated grid cells are found from the Bayesian Optimization approach, the geo-interventions are then a combination of the user constrained grid cells with their optimized variable values (also constrained by user design). Whether the generated intervention is useful is then determined by the AutoML performance, user design constraints on chosen variables and grid cells, and the predicted outcome determined by the optimization metric associated with the target variable target variable y.

3. EXPERIMENTAL EVALUATION

3.1 Experiment

Traffic collisions kill over a million people each year and is one of the leading causes of premature mortality in urban areas (WHO, 2017). There is evidence that infrastructure changes to targeted areas aid in reducing a large share of road traffic collisions in urbanized areas (Noland, 2003). To evaluate the methods mentioned in the previous section, an experiment related to reducing traffic collisions with infrastructure-related geo-interventions in Toronto, Ontario, Canada was conducted.

The experiment involved generating geo-interventions using a variety of publicly available geospatial datasets in Toronto and evaluating the effect of important variables related to infrastructure (detected from the variable importance during the AutoML step) as geo-interventions for reducing road traffic collisions. The experiment also used different grid sizes for the geo-binning step to observe the effects of changes in scale in relation to variable importance and model performance.

3.2 Data

The study area was Toronto, Ontario, Canada. Datasets were downloaded from the City of Toronto Open Data (CTOD) Portal[2] and Toronto Police Service Public Safety (TPS-PS) Data Portal[3]. A total of 21 datasets with 550 variables (columns) and 1,140,927 objects (rows) were used to study geo-interventions for road traffic collisions in Toronto, ON, which were related to transportation (e.g., traffic and red light cameras), infrastructure (e.g., police stations, fire hydrants, and schools), and crime. Table 1 provides the complete list of the 21 datasets[4].

Dataset	Cols.	Rows	Geom.
centrelines	41	70827	Line
collisions	18	499538	Point
traffic	59	226110	Point
autospeedenforcement	6	100	Point
watch_your_speed	14	783	Point
red_light_cams	32	200	Point
police	6	32	Point
ambulance	30	46	Point
fire_hydrants	6	41936	Point
fire_stations	30	92	Point
renewables	41	100	Point
bicycle_parking	20	16998	Point
transit_shelters	24	5852	Point
wayfind	21	330	Point
litter	21	10337	Point
schools	29	1194	Point
childcare	19	1038	Point
art	28	405	Point
culture	32	895	Point
religious	45	1407	Point
crime	28	262707	Point

Table 3. Data used for Toronto traffic collision experiment.

[2] https://open.toronto.ca
[3] https://data.torontopolice.on.ca
[4] https://github.com/rrwen/geogrid-to

Commented [A6]: Can you check the use of tense in this section to make sure either past tense or present tense is used consistently?

Since the past tense is used in the next sections, it may be more consistent to use the same tense here.

Commented [RW7R6]: Done, changed to past tense.

Commented [A3]: Wording?

Commented [A4]: often relates to the time …

Commented [RW5R4]: Done

3.3 Evaluation

To evaluate the experiment, the experiment data in Table 3 was geo-binned to create aggregate variables standardized into several regularized grids of cells (namely sets of 10 by 10, 40 by 40, and 80 by 80 grid of cells). These regularized grids were used as input to two AutoML modelling approaches (TPOT and Auto-Sklearn) each, which created six AutoML regression models (two each for three sets of regularized grid cells). One of the six AutoML models was selected based on the Mean Absolute Error (MAE) to be used for the prediction optimization step.

For the AutoML step, the metric used for guiding the TPOT AutoML models was negative mean squared error, which allows TPOT to minimize the metric or error when building machine-learning pipelines (Wang, Bovik, 2009). The metric used for guiding Auto-Sklearn models was the coefficient of determination R-squared (R^2) (Cameron, Windmeijer, 1997) which allows Auto-Sklearn to measure the performance of each ensembled machine-learning pipeline. Each AutoML model was given one minute of running time to search for the best performing model, where each final model (six total) was compared and measured with the MAE (Chai, Draxler, 2014) to determine the model used for the prediction optimization step. Along with the selected model, grid cells were selected for optimization based on the most important variable. The most important variable was measured by the highest mean permutation importance from the selected AutoML model over 10 randomized runs.

One model was selected out of the six AutoML models to be used for the prediction optimization step. During the prediction optimization step, scenarios were optimized by minimizing the predicted total number of road traffic collisions to generate geo-interventions. Two scenarios were selected based on the variable importance (from permutation importance calculations) to evaluate the simulated effects of the generated geo-interventions. Feasible variables for geo-interventions were manually selected by sorting the variable importance from high to low and observing the list from top to bottom to find three infrastructure-related variables for optimization. In one scenario, grid cells selected for intervention were based on the most important variable, where areas with higher than the average value for the most important variable are selected as cells for optimization. In another scenario, grid cells with lower than average values for each of the three selected feasible variables were chosen as cells for optimization. The differences between the original values and predicted values for each of the three variables for each cell represented the generated geo-interventions, while the difference of total road traffic collisions between the original values and predicted values provided the predicted change from applying the geo-interventions.

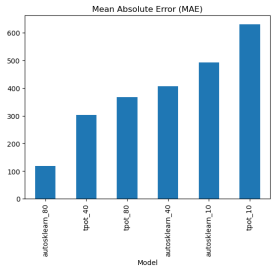

Figure 2. MAE for AutoML models in experiment.

4. RESULTS

The geo-binning step created three sets of regularized grids (10 by 10, 40 by 40, and 80 by 80) for aggregating the 21 datasets in the study area of Toronto, Ontario, Canada. The 550 variables and associated geometries from the data were aggregated into each grid cell to calculate 1417 aggregate variables (columns). The 10 by 10, 40 by 40, and 80 by 80 grids have 100, 1600, and 6400 cells (rows), respectively.

The 10 by 10, 40 by 40, and 80 by 80 grids were fed into TPOT and Auto-Sklearn models to produce six models (two for each grid set). The MAE for each model is shown in Figure 2, where the first part of the model name indicates the AutoML approach and the second part after the underscore indicates the grid size (e.g., TPOT model for a 10 by 10 grid input is *tpot_10*, and Auto-Sklearn for a 80 by 80 grid input is *autosklearn_80*). The Auto-Sklearn AutoML model for a 80 by 80 grid had the best performance (lowest MAE at 117.68), while the TPOT AutoML model for a 10 by 10 grid had the worst performance (highest MAE at 630.08). The most important variable across all models was the total traffic (*traffic_count*) followed by infrastructure related variables for schools, red light cameras, and transit shelters as shown in Figure 3, while the most important variable for the 80 by 80 models was total right-turning westbound car traffic (*traffic_wb_cars_r_sum*) as shown in Figure 4.

The top three most important infrastructure-related variables were schools (*school_counts*), red light cameras (*red_light_cams_count*), and transit shelters (*transit_shelter_counts*). These three variables were fed into the prediction optimization step using the 80 by 80 Auto-Sklearn model (model with lowest MAE). In Scenario 1, grid cells for optimization were filtered to only cells with higher than average traffic for a total of 1860 cells (or parameters for optimization). In Scenario 2, grid cells for optimization were filtered to cells with lower than average schools, red light cameras, and transit shelters respectively for each of the three variables for a total of 17,328 cells. Bayesian optimization was used with 10 iterations to optimize the three most important infrastructure-related variables to minimize the predicted total number of road traffic collisions for each scenario. The generated geo-intervention for Scenario 1 in Figure 5 was predicted to reduce the total number of road traffic collisions by 202,522 collisions from the recorded 429,630 original collisions, while the generated geo-intervention for Scenario 2 was predicted to increase (see Figure 6). The total number of road traffic collisions by 6127 collisions.

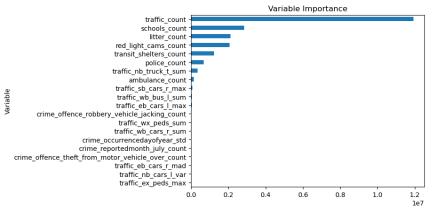

Figure 3. Top 20 important variables across models.

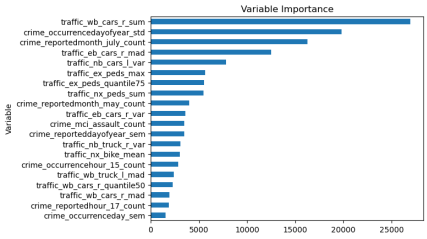

Figure 4. Top 20 important variables from 80 by 80 models.

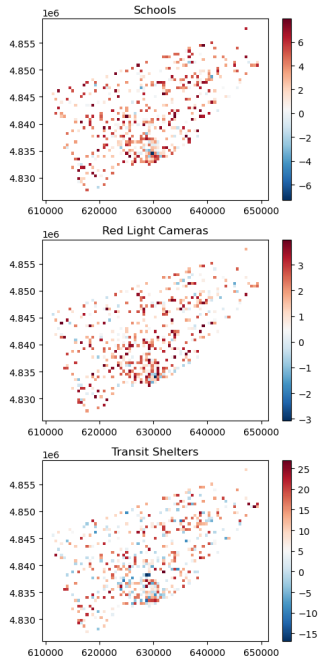

Figure 5. Scenario 1 generated geo-intervention.

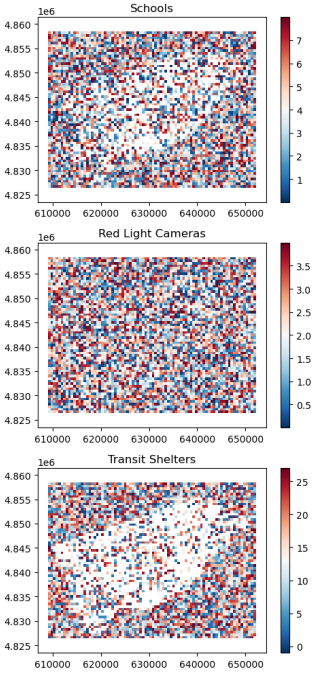

Figure 6. Scenario 2 generated geo-intervention.

5. DISCUSSION

Observing the results reveals that the user-design constraints and decisions have a heavy influence on the prediction optimization step, particularly under low time constraints (10 Bayesian Optimization iterations in this case). Both scenarios used the same three most important infrastructure-related variables (namely *schools_count*, *red_light_cams_count*, and *transit_shelters_count*), but Scenario 2 with the grid cells in scarce infrastructure areas was much noisier and resulted in an increase in total predicted road traffic collisions rather than a decrease. In comparison, Scenario 1 had much lower grid cells with more focused areas of intervention, which resulted in a predicted outcome of approximately halving the total number of collisions. Another note is that the variable importance computed from permutation importance during the AutoML modelling step helped focus the optimization process on particularly influential variables and guided the grid cell filtering constraints made by the user. This is especially helpful in cases where there are a large number of variables and a large number of grid cells. The choice of grid cell dimensions during the geo-binning step is also important, as lower resolution grids (e.g., 10 by 10 grid) resulted in AutoML models with much higher MAE scores of above 400 versus higher resolution grids (e.g., 80 by 80) resulted in models with MAE scores below 400.

Limitations to this research includes subjectivity in user design constraints, scalability, and interactivity. Although the geo-binning, AutoML, and the optimization process are largely

automatic, user design constraints are reliant on the user and empirical experimentation, which heavily influence the generated geo-intervention outcomes. Related to user design subjectivity are scalability and interactivity. Depending on the data size and resources available, the major restriction becomes time needed for building models, calculating variable importance, and optimizing predictions, which require a more iterative approach to explore variables and different user design constraints to find appropriate running times and optimization constraints (Jiang, Liu, and Chen, 2019). Future work may be studying distributed computing related to the AutoML and optimization processes, and developing an interactive graphical interface for exploring user design constraints and generated geo-interventions that balance efficiency with performance.

6. CONCLUSION

This research proposed an approach for a SDSS that integrated AutoML and HPO with GIS to leverage modern advancements in computing power and big data availability. The geo-intervention generation approach involved three steps of geo-binning, AutoML modelling, and prediction optimization. Geo-binning standardized tabular point, line, and polygon geospatial data into regularized grid cells with aggregated variables. AutoML then takes these regularized grids and modelled a target variable representing the geo-intervention outcomes desired. This step also produced variable importance that helped provide insight on the effects of important variables relative to the AutoML model, which helped guide the user to define better design constraints in the next step. Finally, the prediction optimization step used the AutoML model and user design constraints (selection of important variables and grid cells to optimize) to generate precision geo-interventions in which the predicted outcome is desirable. The geo-intervention generation approach was evaluated with an experiment focused on reducing road traffic collisions with infrastructure changes in Toronto, Ontario, Canada. A total of 22 datasets with 1,140,927 geospatial objects and 550 variables were geo-binned into several grid cells of 10 by 10, 40 by 40, and 100 by 100, where each grid had a total of 1417 aggregate variables. Six AutoML models were produced (two for each set of three grids). The top three most important infrastructure variables from the best performing model (80 by 80 Auto-Sklearn model) with an MAE of 117.68 were used in the prediction optimization step to generate two geo-interventions. The geo-intervention for Scenario 1 was targeted in high traffic areas and predicted to reduce the total road traffic collisions by approximately a half, while the geo-intervention for Scenario 2 was noisy and more randomly spread out and was predicted increased the total road traffic collisions by over 6000. From experimental observations, limitations included subjectivity in user constraints, scalability, and interactivity. Although precision geo-interventions were generated with AutoML and cover a much larger range of considerations and possibilities from big data and high computation, future work to improve efficiency in modelling/optimization processes and user exploration is essential to build user trust, and to help users explore more optimal geo-interventions.

REFERENCES

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

Cameron, A. C., Windmeijer, F. A., 1997: An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of econometrics*, 77(2), 329-342.

Chai, T., Draxler, R. R., 2014: Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development*, 7(3), 1247-1250.

Feurer, M., Hutter, F., 2019: Hyperparameter optimization. *Automated machine learning*. Springer, Cham, 3-33.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F., 2015: Efficient and robust automated machine learning. *Advances in neural information processing systems*, 28.

He, X., Zhao, K., Chu, X, 2021: AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212, 106622.

Jiang, L., Liu, S., & Chen, C. (2019). Recent research advances on interactive machine learning. Journal of Visualization, 22(2), 401-417.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., ..., Cheng, T, 2016: Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, 115, 119-133.

Malczewski, J., Rinner, C., 2015: Introduction to GIS-mcda. In Multicriteria decision analysis in geographic information science. *Springer*, Berlin, Heidelberg, 23-54.

Mehaffy, M. W., 2008: Generative methods in urban design: a progress assessment. *Journal of Urbanism*, 1(1), 57-75.

Noland, R. B., 2003: Traffic fatalities and injuries: the effect of changes in infrastructure and other trends. *Accident Analysis & Prevention*, 35(4), 599-611.

Olson, R. S., Bartley, N., Urbanowicz , R. J., Moore, J. H., 2016: Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. *Proceedings of GECCO 2016*, 485-492.

Snoek, J., Larochelle, H., Adams, R. P., 2012: Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

Wang, Z., Bovik, A. C., 2009: Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1), 98-117.

WHO, 2017. Save lives: a road safety technical package. https://www.who.int/publications/i/item/save-lives-a-road-safety-technical-package (19 Sept 2022).

Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H., 2019: Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, 17(1), 26-40.

Commented [A8]: Use of tense?

Commented [A9]: What do you mean about “due to experimental observations”?