Towards a generalizable Urban Heat Island assessment tool using data-driven models

Author: Marco Dikkers

Abstract:

Cities all over the world are increasingly exposed to the Urban Heat Island (UHI) phenomenon as a result of global warming and urbanization. Data-driven models have gained popularity over the recent years and are used to analyze UHI intensity in different urban settings. These models specifically address the need for an easy-to-use assessment tool that can incorporate UHI concerns in the decision-making of urban planning. Multiple data-driven models that employ location specific data have been developed and successfully validated for Land Surface Temperature (LST) estimation in cities. It would be of great value to be able to reuse the trained UHI models for multiple cities, because this saves the effort to train the model with location specific data for every city in advance. Yet, previous research provides very little insight into the extent to which such data-driven models are generalizable for cities that are different in size, population, and regional climate. This research aims to conduct a comprehensive generalizability study of a Random Forest (RF) regression approach in relation to different levels of similarities between the urban characteristics of the cities used in this research. To this end, five different cities from three different countries were selected to cover a diverse range of (dis-)similarities in urban contexts. The individual models that were developed for each city are shown to be accurate in LST estimation in the cities for which they were trained. However, external cross-validation of the model to data from other urban contexts reveals that the proposed data-driven models have very low generalization capabilities, regardless of the observed (dis-)similarities between the cities. It was concluded that small changes in the feature properties can result in significant variation in the UHI behavior, and therefore the generalization is deficient. The results of this research show that the emergence of UHI is very context-specific, and that implies that implementation of standardized/universal mitigation strategies across cities world wide may be inappropriate. Instead, urban planners should address UHI in a location-specific manner, considering the local UHI mechanism for any given city.

Keywords: Urban Heat Island, Data-driven modeling, Random Forest, Urban planning, Generalizability

1 INTRODUCTION

The construction sector is currently facing several major challenges as a result of climate change and the trend that is observable in the growing world population. Urban areas are expanding horizontally and vertically as more people converge towards large cities to settle. The United Nations anticipates the urban areas (i.e., cities) to be inhabited by 68% of the world's population by 2050 (United Nations, 2018). As a result, urban planners are left with the challenge of accommodating sustainable, safe and livable environments in cities, as those are expected to grow considerably in the near future. On top of urbanization, environmental impacts will bring additional issues to the drawing board for city developers in the coming years. The properties of building materials that are comonly used in urban areas (i.e., colors, surface roughness, thermal conductivity) lead to increased storage of solar radiation in cities compared to their rural environments. Furthermore, anthropogenic heat released from buildings due to human activities and the lacking availability of moisture to evaporate from hard surfaces also contribute to temperature rise in cities (Ahmed Memon et al., 2008; Pena Acosta et al., 2021a). The phenomenon of heat generation in urban areas is commonly known as the Urban Heat Island (UHI) effect. Human health, living comfort, and the local economy suffer from extremely high temperatures in urban areas that are prone to the UHI (Akbari et al., 2016). Among the most vulnerable segments of society are, for example, elderly or very young people (Akbari et al., 2016), people living in low-income housing (Sakka et al., 2012), but also people that are performing long lasting physical work in warm environments, like for example construction workers that have died from heat-related illnesses (Acharya et al., 2018).

Needless to say, the UHI effect plays an important role in the design of urban environments. Design decisions in urban planning are to a great degree made at street level, stressing the necessity to provide comprehension about UHI intensity at this scale. But in practice, UHI analyses are not widely considered by urban planners due to the complexity of UHI modelling. Conventionally,

architects and sustainability engineers use physics-based simulation methods to model the UHI effect at micro- (i.e., individual buildings or street segments) and local-scale (i.e., neighbourhoods). These physics-based approaches rely on governing fluid dynamics principles, such as thermal convection, solar radiation exchange, and air ventilation around buildings (Mirzaei, 2015). EnergyPlus, Envi-met, and Urban Weather Generator are some examples of simulation software packages that are commonly used among architects and sustainability engineers to evaluate UHI. Among urban planners however, such tools are hardly used in the decision-making process. One major drawback of physics-based modeling is that the scale to which it can be applied is limited due to the computational capacity that is required for the calculations (Mirzaei, 2015). They require extensive representations of the building geometries and properties of building materials, which results in that similuations are often simplified (i.e., reduced to one building or urban canyon). Furthermore, developing these models can be time-consuming, and simulation results can become too complex to be interpreted and implemented in the decision-making of urban design by average city planners.

Meanwhile, data-driven models have gained more popularity for UHI assessment at micro-scale over the last years. The increasing availability of urban data and data-driven models are used to find correlations between simplistic urban characteristics (i.e., low level of detail in the properties) and heat generation, without the need of modeling the heat exchange process, which involves numerous parameters and properties. One advantage of data-driven modeling is that the end user does not necessarily need full understanding of how output variables are predicted using the given independent input variables. Many Machine-Learning (ML) applications are known for analyzing buildings' energy performances and other UHI-related issues for specific urban environments (Gobakis et al., 2011; Pena Acosta et al., 2021a, 2021b; Wu et al., 2019; Zhang et al., 2019). The recent work of the authors (Pena Acosta et al., 2021b) implemented both a Random Forest (RF) and Decision Tree (DT) regression approach to make a distinction between five UHI intensity

levels at street level, using publicly available datasets carrying geospatial information obtained from Geographic Information System (GIS) models. The DT regressor was applied to the city of Montreal, Canada, and performed very efficiently with an accuracy of 93%. However, a common question for data-driven models is to what extent they are generalizable. In this case, the question was raised how accurate the ML model can estimate UHI intensity in case it was trained employing data from one single city, and asked to predict temperatures in a city that is different in size, environmental climate, urban morphologies (e.g., building geometry, average building height to street width ratio (H/W), built-up density, water bodies, and vegetation) and socio-economic characteristics (e.g., land use, population density, and traffic flow).

A generalizable data-driven predictor of UHI intensity could be of great value for urban planners, as it would yield high accuracies in UHI prediction at micro-scale for any given circumstances, without the necessity to employ location specific data to train the model in advance. Furthermore, such models could substitute physics-based models that are known to be complex and time-consuming in practice. Nevertheless, a recent study (Pena Acosta et al., 2021b) demonstrated that in case the model is asked to predict temperatures outside the context of the training dataset in a different urban environment, the algorithms perform poorly for that particular case. This suggests that the generalizability of the model is limited. However, to the best of the author's knowledge, the coverage of current research on the domain of generalizable data-driven UHI assessment is insufficient as it is not yet applied to urban areas that are more similar in the core characteristics that affect UHI intensity in cities. Furthermore, whilst several studies investigated the influence of different factors on the UHI effect, quantification of the importance of these factors is sparce in the related literature. (Sangiorgio et al., 2020) quantified the contribution of a number of factors to UHI intensity for the first time. The results of their study suggest that presence of vegetation and the urban albedo (i.e., capacity of urban surfaces to reflect solar radiation) are the most important factors for UHI, followed by population density, street widths, canyon orientation and

building height. These empirical results were obtained from case studies in 41 European cities. However, supplementary researches that support these findings are lacking. Moreover, the inequality in the degree to which different factors contribute to the UHI effect in different urban contexts is not fully understood.

In this research, a comprehensive generalizability study for a data-driven UHI assessment tool is conducted, considering five different cities that have varied levels of (dis-)similarities in the core characteristics that correlate to the UHI effect. This aims to better understand the importance of factors influencing UHI in different urban contexts, and the extent to which the proposed data-driven modelling approach is generalizable for cities from different environments, sizes, and populations. This paper is structured as follows: The next section describes the methodology for assessing the generalizability of the proposed method in different scenarios. Thereafter, the results of the generalizability study are presented. The paper closes with a discussion and conclusions drawn from the research results presented in this study.

2 RESEARCH METHODOLOGY

Figure 1 shows a schematic representation of the research framework, including data collection, data processing, and data analysis. The generalizability of the data-driven model for different urban contexts is assessed by means of a combination between two measures, which are the most prominent outputs of the framework. The first is the performance of the ML models resulting from external cross-validation (i.e., applying the optimized model of each city to samples from the other four cities, that are different in urban context). The second is the pairwise quantification of the (dis-)similarity in urban context that exists between the cities under consideration. This section describes the research methodology, following the elements in the research framework of figure 1.

2.1 Data Collection

2.1.1 City selection

For this study, five different cities were selected of which publicly available geospatial data was gathered. The criteria to select the cities in this research should be based on the ability to measure similarities and dissimilarities in the urban characteristics that influence UHI. It is hypothesized that the model will be generalizable for cities that are similar in the factors that have the most impact on UHI intensity at street level. Yet, differences in these characteristics are reflected in the distributions (i.e., mean, variance, and the shape of the distribution) of the features that are used in the model, and are therefore indistinguishable before the data for the selected cities has been collected and processed. Accordingly, cities were selected based on their proximity and (dis-)similarities in size, population and climatic environment. The similarities derived from the feature distributions are quantified after the data has been collected and processed, as will be explained in subsection 2.3.3. Table 1 lists the five cities that were selected for this research. The regional climate, city size, and population are presented as indicators for the environmental, urban morphological and socio-economical factors respectively.

2.1.2 Feature selection and data availability

The data used in this study are available through open data sources, mainly operated by local governmental institutions. Hence, the cities in this study do not share the same data sources. The characteristics and structure of the data however, remains similar for all cities in this study in order to allow comparison between UHI assessment in different scenarios. The availability of the data through open sources has been a boundary condition for the feature selection in this research. The data that were gathered as input for the ML model carry information about 11 features (i.e., explanatory variables) that influence the UHI intensity, and the land surface temperature (LST) as response variable. The inclusion of most features (building geometries, vegetation, waterbodies, H/W ratio, population density, and land use) were adopted from a proposed data-driven model

(Pena Acosta et al., 2021b) that has shown to be accurate in estimating LST at street level. The ground surface elevation, retrieved from Digital Elevation Models (DEM), is added to the ML model as it known to have an significant impact on UHI intensity (Wu et al., 2019). Table 2 lists the features and the corresponding publicly available sources of the data that are used for the cities that were analyzed in this research. The LST data for all cities are derived from Landsat 8 satellite images (Landsat Data Access | U.S. Geological Survey), employing multiple images taken in the summers between 2019 and 2021. Only satellite images were used that have a proportion of cloud cover less than 30% to reduce the noise in the temperature data. After establishing the availability of the data required for all cities, the data will be collected and processed afterwards.

2.2 Data Processing

The processing of the data is carried out in ArcGIS mapping software. The general workflow for data processing is consistent with the proposed approach (Pena Acosta et al., 2021b). The following subsections provide a brief elaboration on the data processing approach.

2.2.1 Specify Street Buffers

The urban feature data and LST data that have been collected for each city has to be rearranged at street level, such that each observation reflects and distinguishes the local socio-economical and urban morphological characteristics, and the LST of that particular street segment. To this end, a buffer is created around the centre line of each street segment, used to capture and group the feature characteristics for each observation into one unit. The buffer distance was set at 15 meter in the proposed model (Pena Acosta et al., 2021b), and kept equal in this study. Figure 2 shows an example of a buffer area that is computed around a centre line of a street segment.

2.2.2 Compute Feature Values

The features that are used to develop the databases take different forms, and require different computations, as described hereafter. The *street widths* for New York were directly available in

the data source that is listed in table 2. For the other cities, the widths of the streets were not available as a feature in the publicly available data. Hence, the street widths were obtained by calculating the average closest distance along the street centre line to the outside boundary of the roadbed. The densities per street segment (i.e., *building density*, *vegetation density*, and *water density*) were computed by the taking the proportion of the buffer area that is overlapping with either buildings, vegetation, or water respectively. Figure 3 shows an example of how the overlapping areas of buildings, vegetation, and water are isolated in a street buffer, using the ArcGIS software. The different land use descriptions have been labeled according to eight categories listed in table 3. Subsequently, the *predominant land use* is represented by the land use category of which the proportion is most dominant within the buffer area, as illustrated in figure 4. The population data for all cities are composed of grids that return the number of people per cell. The *mean population* per street section is computed by the weighted average (i.e., taking the proportions of cells that are within the buffer into account) of the population in each cell, as illustrated in figure 5. The *mean building height* is also derived from the weighted average within the buffer. The *elevation* is obtained by taking the average value of the DEM within the buffer area.

2.2.3 Calculate ΔLST

The intensity of UHI is generally expressed in terms of the LST differential (ΔLST) in the urban area, compared to its rural environment (Oke, 1973). The LST is estimated using Landsat 8 thermal bands using (USGS Landsat Level-1 Data Product | U.S. Geological Survey). The satellite images were processed according to the NDVI algorithm (Sobrino et al., 2004). The availability of useful satellite images between 2019 and 2021 per city varied from a maximum of nine for the cities of Apeldoorn, Rotterdam, New York, and Montreal, to a minimum of five for the city of Enschede. The value for ΔLST in each observation is calculated by the difference between the average LST within the created buffer, and the averaged reference LST measured at three independent locations

in the rural environment of the city. Figure 6 shows a LST map of the city of Enschede after processing the Landsat 8 images, and the locations of three reference points in the rural environment of the city. Worth mentioning is that, different from other comparable data-driven UHI assessment studies, this research aims to investigate the generalizability of the model for different urban contexts. Many studies on the causes and mitigations strategies of UHI point out that the order of magnitude of UHI intensity is sensitive to the local environment, and in particular the local baseline temperature. This makes comparison of ΔLST between cities from different environments inappropriate. For instance, the average observed LST for New York City was 37.0 °C, while it was 24.3 °C for the city of Apeldoorn. This difference between the observed urban environments will bring noice to the training datasets that are used for both cities. Accordingly, the average percentage difference ΔLST [%] for each observation, relative to the reference LST in the local rural environment, is used as responsive variable. Thus, the magnitude of ΔLST per city, in terms of percentage is used to account for environmental differences between cities. The ΔLST [%] for each observation is averaged using the created buffer areas, resulting in 20% increase in LST on average for Apeldoorn, and 17% for New York City, as an example. The computed values for the features and the ΔLST [%] of each observation are combined, and the data is cleaned (i.e., duplicate removal and fixing structural errors), resulting in five datasets of which the general structure is shown in table 4.

2.3 Data Analysis

2.3.1 Data selection

Due to the differences in size of the observed cities, the available data populations per urban context will vary. Generally, the performance of ML models improves as the data population used for training increases. However, due to the need for comparison between the models that are developed for this study, a consistent number of data instances is used for all cities to develop the best-performing model, and perform external cross-validation. The populations that are used for

each of the cities are equal to the urban context that yields the smallest data population in total. For the remaining cities, the data instances are randomly selected from the total population.

2.3.2 Individual RF regressors

An RF regression approach is implemented, as random forests are known to be less prone to overfitting than DTs due to the use of multiple randomly generated trees. This makes RF more suited for generalization problems in general (Breiman, 2001; Lan et al., 2020). The implementation of the RF algorithm was carried out in the Python distribution platform, utilizing the scikit-learn library (Pedregosa et al., 2011). The selected data populations are randomly split into subsets with a 70:30 training and testing ratio. Two methods for hyperparameter tuning are used from the scikit-learn library. First, RandomizedSearchCV is applied to narrow down the range for each hyperparameter (i.e., number of estimators, minimum samples per split, min samples to reach a leaf, maximum depth, bootstrapping). Subsequently, GridSearchCV is used to obtain the hyperparameter settings of the best-performing models. The performance of the RF regressor is estimated by means of the Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE) in the prediction of response variable ΔLST [%]. For each model, the goodness of fit is estimated by means of R-squared (R^2).

2.3.3 Similarity Index

To better understand the generalizability of the model for different urban context, the (dis-)similarities in the feature properties between each of the models are studied. For all combinations of two cities a similarity index (SI) is developed that represents the degree of (dis-)similarity that exists between the two cities. First, the relative contribution of the features that are used in the model is investigated. To this end, the feature importances are extracted from the best performing RF regressor of each city, using the scikit-learn library. Subsequently, the dissimilarities in the distributions per feature are examined. A two-sample Kolmogorov-Smirnov (KS) test is used to quantify the distance between two empirical distributions, considering the KS-statistic (D) as a

measure for the dissimilarity. The distance between the given distributions of feature *F(x)* of two cities *y* and *z* is obtained following equation (1):

$$(1)\ D_{yz} = \sup_x \left| F_y(x) - F_z(x) \right|$$

Where D_{yz} is the KS-statistic (i.e., distance) corresponding to cities *y* and *z*, and sup_x is the supremum (i.e., largest absolute difference) of the cumulative functions $F_y(x)$ and $F_z(x)$. The feature importances (FI) of the best performing model of city *y* operate as a weighing factor to calculate dissimilarities. The SI is calculated by taking the sum of all feature distances, multiplied by the corresponding feature importances. The calculation of the SI is expressed in following equation (2):

$$(2)\ SI_{yz} = \sum_i^n D_i * FI_i + D_{i+1} * FI_{i+1} \ldots + D_n * FI_n$$

Where SI_{yz} is the similarity index for the best performing model of city *y* cross-validated to city *z*. D_i is the KS-statistic for feature *i* (obtained from equation (1)), FI_i is the feature impotance considering feature *i*, and *n* is the number of features used in the RF regressor. Since external cross-validation between two cities is performed in two ways (i.e., city *y* to city *z*, and city *z* to city *y*), the SI in each direction is different because the feature importances of city *y* are not equal to those of city *z*. The structure of table 5 is used to calculate the SIs for all scenarios.

2.3.4 External cross-validation

The best performing model for each city in terms of R^2 is asked to predict ΔLST [%] for the unseen data of the remaining four cities, resulting in 20 external cross-validations in total. In the next section, the generalizability of the RF regressor is assessed for different scenario's, considering the varying levels (dis)-similarities, expressed in the SIs.

3 STUDY RESULTS

This section presents the results of the generalizability study, arising from the research framework, depicted in figure 1. An elaborate discussion on the study results is provided in the upfollowing

section. The collection and processing of the data resulted in five independent datasets, structured according to table 4. Due to the differences in size between the observed urban areas, the datasets vary in population size (see table 6 for data populations per city). The city of Montreal yields the smallest number of observations (slightly over 5000). In contrast, the data for New York City adds up to a total of 83.000 instances, making this dataset the largest. The histograms of the data in terms of ΔLST [%] for all cities are presented in figures 7 – 11.

The dataset of each city that is used for further analysis is reduced to a total number of approximately 5000, selected from the total data populations. Random selection however, results in unbalanced datasets in terms of ΔLST [%], especially for the larger populations like New York and Rotterdam. Since only a small proportion of the total population is selected for these cities (5000 out of 83.000 and 15.000 respectively), it is more likely that instances close to the mean are selected, as there are considerably more observations in this range compared to the lower and higher values of ΔLST [%]. As a result, the data instances with relatively high and low UHI intensities are overlooked. Likewise, the RF regressor gains the most information from a more varied dataset. Particularly the higher and lower UHI intensities are of major interest for city developers, in order to evaluate mitigation strategies. Therefore, the 5000 data instances are picked from the total population, such that the range and variance in terms of ΔLST [%] is maximized for each city. Figures 12 – 16 show the histograms of the datasets that were used for training and testing of the models.

The individual models for each city are trained using the best hyperparameter settings (i.e., yielding the highest value for R^2). Table 7 shows the feature importances per city, derived from the best-performing RF regressors. A sixth row was added to show the average of the feature importances that were derived from the best-performing models, and a seventh row that shows the feature importance, based on a model trained on a mixed dataset (i.e., containing the 5000 data instances of each city). Subsequently, the SIs for the different scenarios are calculated, considering

the KS-statistic for the feature distance, multiplied by the feature importance retrieved from the optimized model. Table 8 shows the SIs for all scenarios that were analyzed in this study. The values in the diagonal of the grid are equal to 1.00, since the cities used for training and testing in these scenarios constitute the exact same urban context. Table 9 summarizes the performance of the RF regressor in terms of MAE, R^2 and MAPE for different scenarios. The metrics for the best-performing models per city are positioned in the diagonal of the grid. Figures 17 – 21 show the scatterplots, resulting from the best performing models and external cross-validation for each city. The correlation between tables 8 (Similarity Indeces) and 9 (model performance) is further analyzed to divulge the generalizability of the proposed model, in relation to the (dis-)similarities between the cities. To this end, a Pearson's correlation test is performed, summarized in table 10. Because the external cross-validation of all models signifficantly reduced performance of the models, as will be explained in the discussion section, five additional scenarios were analyzed in which the model was trained on a mixed dataset containing the 5000 data instances from four cities, and tested on the fifth city of which the data was excluded from the training dataset. The reasoning behind this strategy is that all individual models are only able to capture location specific UHI behavior, and therefore, it is assumed that the regression models fail to generalize for the other cities. The advantage of mixing the datasets with feature vectors from different cities is twofold. First to mention, the model is trained on a larger dataset, which generally improves the accuracy and generalizability of RF regressors. Second, a mixed dataset captures a more averaged relationship between UHI intensity and the feature values of different cities, reducing the chance of overfitting the training data obtained from one specific urban context. In the five additional scenarios, the model is trained on a mixed dataset containing approximately 20.000 data instances from four out of the five cities, and external cross-validated on the 5000 data instances from the city that was not included in the mixed data population. Figures 22 – 26 show the scatterplots and

performance metrics of the five scenarios. In the following section, the results of the generalizability study are discussed.

4 DISCUSSION AND CONCLUSIONS

Recent research has shown that the proposed data-driven method to analyze UHI intensity at micro-scale could be a reliable substitute for conventional physics-based models, which are known to be too complex and time consuming for average city planners. The use of simplistic feature vectors that are easily obtained through open data sources, and the interpretability of the models make this an easy-to-use tool for the assessment of UHI. This section discusses the results of the conducted generalizability study for the proposed model.

All individual ML models show good predictability within the scope of the training datasets that were used to develop the RF regressors for each city. The data for Enschede, Apeldoorn, and Montreal fit the regression models best, indicated with an R^2-value around 0.7. The R^2-values for Rotterdam and New York are 0.56 and 0.63 respectively. This difference could be explained by the fact that a smaller proportion of the total data population was used for the cities of Rotterdam and New York, and UHI behaviour within the scope of larger cities may vary more in different neighbourhoods, compared to smaller cities. The model of Montreal performs best with MAE equal to 0.02, which is equal to an everage error of 0.49 ℃ (0.02 ΔLST [%] error relative to the reference temperature which was 24.5 ℃ for Montreal), and MAPE equal to 0.13, as shown in figure 21(a). The model of Rotterdam performs worst with an MAE equal to 0.07, which is an average error of 1.68 ℃, and MAPE of 4.45, shown in figure 19(a).

From the best performing models, the feature importances were derived. On average, the ground surface elevation, vegetation density, and landuse are the most important features, followed by the population density, building density, street width, and maximum building height. These outcomes are quite consistent with the findings of comparable research, that also marked vegetation, population, street widths, and building height as important factors for UHI (Sangiorgio et al.,

2020). Noteworthy, not all features that were analyzed are exactly the same in both studies, and these observations are the result of average UHI behaviour and do not necessarily represent all individual cities. The feature importances that were derived from the mixed dataset are more or less consistent with the average feature importances of all the individual models. When looking at the feature importances of the models (table 7), the individual values hover around the average feature importance. The maximum deviation from the averaged values are the feature importances of vegetation density and population density for the city of Enschede (0.137 and 0.115 higher than the mean respectively). All other values are within 0.1 range of the average feature importances.

In this research, (dis-)similarities in the most important urban characteristics between the observed cities were quantified. As indicated by tables 1 and 8, the resulting SIs are very little related to the geographic proximity, and the differences in size and population of the cities in this study. However, while the SIs are the result of the differences in feature importance and the feature distributions, they still might help in explaining the generalizability potential of the model for different cities.

The individual models for each of the cities show good predictability within the context of the training data. The model of Rotterdam performs worst, where R^2 is 0.56 and MAPE is 4.45, but still shows good predictive capability. This confirms the validity of the model in case it is applied to the context of the training data. However, for all of the 20 scenarios in which the trained model was cross-validated to data from another city, the performance decreases drastically. The wide spreaded observations in the scatterplots and low R^2 values in figures 17 (b, c, d, and e) – 21 (b, c, d, and e) are indicators for low generalizability of the model. None of the scenarios that were analyzed show reasonable predictive capacity outside the scope of the training data. The correlation between the SI and the cross-validated models was further investigated. As can be seen in table 10, the high P values and low values for the correlation coefficient (r) for all three performance metrics suggest a weak correlation between the SIs and the model performances. This

weak correlation stresses the strong dependency of the ML model on the specificities of the urban context (i.e., small variations in the model's feature properties result in significant variation in the predicted UHI intensity). Although the differences in feature importances are relatively small, as shown in table 7, the model is not generalizable for the cities other than the one that was used for training.

In five additional scenarios, the model was trained on a mixed dataset and cross-validated on data instances from a city that was excluded from the training data. As can be seen in figures 22 – 26, the mixed models performed slightly worse than the individual models that were developed for each city. The worst performing model concerned the one that excluded Enschede from the mixed data, and performed with an R^2-value of 0.61 and MAPE of 3.44. Yet, when the mixed models were cross-validated to the unseen data of the excluded city, again the performance dropped significantly. This, in addition to the other scenarios that were analyzed in this study, indicate that the models are too sensitive to small differences in urban context, eventhough the variety and quantity of the training data was expanded.

From the conducted research it can be concluded that model has low generalizability for cities from different contexts. It appears that UHI behavior is very sensitive to the local environment, and that the magnitude of the factors contributing to UHI may vary for different environments. This finding indicates that world wide universal strategies to mitigate UHI intensity may have different outcomes in different urban contexts, and therefore, the implementation of a standardized 'one-size-fits-all' strategy may be inappropriate, or at least suboptimal. Instead, the UHI phenomenon should be studied and addressed in a context-specific manner, considering the local driving factors of UHI for any given urban context. The proposed data-driven modeling approach shows great potential in revealing the UHI mechanism for any city and it is fairly easy to implement, in spite of the differences in how the required data are structured and stored by different governmental institutions. This indicates the strong possibility and adding value of using this

approach in urban planning decision-making, as it provides a proper substitute conventional assessment methods.

5 REFERENCES

1 foot Digital Elevation Model (DEM) | NYC Open Data. (n.d.). Retrieved October 24, 2022, from https://data.cityofnewyork.us/City-Government/1-foot-Digital-Elevation-Model-DEM-/dpc8-z3jc/data?no_mobile=true

3D Basisvoorziening - PDOK. (n.d.). Retrieved October 24, 2022, from https://www.pdok.nl/3d-basisvoorziening

Acharya, P., Boggess, B., & Zhang, K. (2018). *Assessing Heat Stress and Health among Construction Workers in a Changing Climate: A Review*. https://doi.org/10.3390/ijerph15020247

Ahmed Memon, R., Leung, D. Y., & Chunho, L. (2008). A review on the generation, determination and mitigation of Urban Heat Island. *Journal of Environmental Sciences*, *20*, 120–128.

Akbari, H., Cartalis, C., Kolokotsa, D., Muscio, A., Pisello, A. L., Rossi, F., Santamouris, M., Synnefa, A., Wong, N. H., & Zinzi, M. (2016). Local climate change and urban heat island mitigation techniques – the state of the art. *Journal of Civil Engineering and Management*, *22*(1), 1–16. https://doi.org/10.3846/13923730.2015.1111934

Breiman, L. (2001). Random Forests. *Machine Learning 2001 45:1*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Centraal Bureau Statistiek. (n.d.). *Kaart van 100 meter bij 100 meter met statistieken*. Retrieved October 24, 2022, from https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/kaart-van-100-meter-bij-100-meter-met-statistieken

DoITT. (n.d.). *Understanding the 2017 LiDAR Capture*. Retrieved October 24, 2022, from https://maps.nyc.gov/lidar/2017/

Geo services - PDOK. (n.d.). Retrieved October 24, 2022, from https://www.pdok.nl/geo-services/-/article/cbs-bestand-bodemgebruik

Gobakis, K., Kolokotsa, D., Synnefa, A., Saliari, M., Giannopoulou, K., & Santamouris, M. (2011). Development of a model for urban heat island prediction using neural network techniques. *Sustainable Cities and Society*, *1*(2), 104–115. https://doi.org/10.1016/J.SCS.2011.05.001

Lan, T., Hu, H., Jiang, C., Yang, G., & Zhao, Z. (2020). A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification. *Advances in Space Research*, *65*(8), 2052–2061. https://doi.org/10.1016/J.ASR.2020.01.036

Landsat Data Access | U.S. Geological Survey. (n.d.). Retrieved October 14, 2022, from https://www.usgs.gov/landsat-missions/landsat-data-access

Mirzaei, P. A. (2015). Recent challenges in modeling of urban heat island. *Sustainable Cities and Society*, *19*, 200–206. https://doi.org/10.1016/J.SCS.2015.04.001

Montreal Open Data Portal. (n.d.). *Québec - Données Québec*. Retrieved January 11, 2023, from https://www.donneesquebec.ca/organisation/ville-de-quebec/

NYC 3D Model | NYC Open Data. (n.d.). *NYC 3D Model Download*. Retrieved January 11, 2023, from https://www.nyc.gov/site/planning/data-maps/open-data/dwn-nyc-3d-model-download.page

NYC Street Centerline (CSCL) | NYC Open Data. (n.d.). Retrieved October 24, 2022, from https://data.cityofnewyork.us/City-Government/NYC-Street-Centerline-CSCL-/exjm-f27b

Oke, T. R. (1973). City size and the urban heat island. *Atmospheric Environment (1967)*, *7*(8), 769–779. https://doi.org/10.1016/0004-6981(73)90140-6

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot and Édouardand, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu

Perrot. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-learn.sourceforge.net.

Pena Acosta, M., Vahdatikhaki, F., Santos, J., Hammad, A., & Dorée, A. G. (2021a). How to bring UHI to the urban planning table? A data-driven modeling approach. *Sustainable Cities and Society*, *71*. https://doi.org/10.1016/J.SCS.2021.102948

Pena Acosta, M., Vahdatikhaki, F., Santos, J., Hammad, A., & Dorée, A. G. (2021b). A generalizability analysis of a data-driven method for the Urban Heat Island phenomenon assessment. *Proceedings of the International Symposium on Automation and Robotics in Construction*, *2021-November*, 73–80. https://doi.org/10.22260/ISARC2021/0012

PLUTO - NYC DCP . (n.d.). Retrieved October 24, 2022, from https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page

Sakka, A., Santamouris, M., Livada, I., Nicol, F., & Wilson, M. (2012). On the thermal performance of low income housing during heat waves. *Energy and Buildings*, *49*, 69–77. https://doi.org/10.1016/j.enbuild.2012.01.023

Sangiorgio, V., Fiorito, F., & Santamouris, M. (2020). Development of a holistic urban heat island evaluation methodology. *Scientific Reports 2020 10:1*, *10*(1), 1–13. https://doi.org/10.1038/s41598-020-75018-4

Sobrino, J. A., Jiménez-Muñoz, J. C., & Paolini, L. (2004). Land surface temperature retrieval from LANDSAT TM 5. *Remote Sensing of Environment*, *90*(4), 434–440. https://doi.org/10.1016/J.RSE.2004.02.003

United Nations. (2019). *World Urbanization Prospects The 2018 Revision*. https://population.un.org/wup/publications/Files/WUP2018-Report.pdf

Using the USGS Landsat Level-1 Data Product | U.S. Geological Survey. (n.d.). Retrieved October 18, 2022, from https://www.usgs.gov/landsat-missions/using-usgs-landsat-level-1-data-product

WorldPop - Population Counts. (n.d.). Retrieved October 24, 2022, from https://hub.worldpop.org/geodata/summary?id=49727

Wu, X., Zhang, L., & Zang, S. (2019). Examining seasonal effect of urban heat island in a coastal city. *PLOS ONE*, *14*(6), e0217850. https://doi.org/10.1371/JOURNAL.PONE.0217850

Zhang, K., Dong, X., Liu, Z., Gao, W., Hu, Z., & Wu, G. (2019). Quantifying the Effects of Urban Form on Land Surface Temperature in Subtropical High-Density Urban Areas Using Machine Learning. *Remote Sensing*, *11*(8), 959. https://doi.org/10.3390/RS11080959

FIGURES

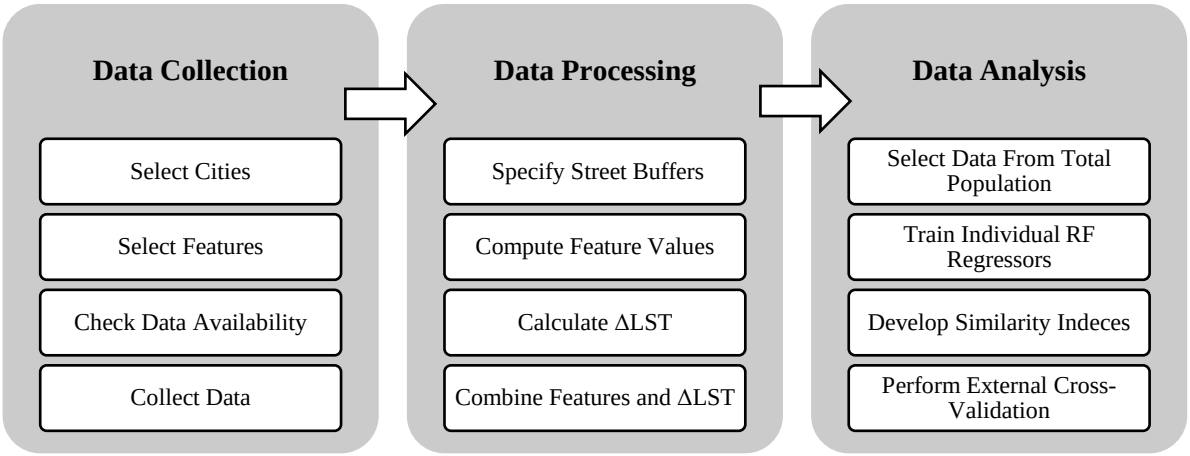


Figure 1. Conceptual research framework

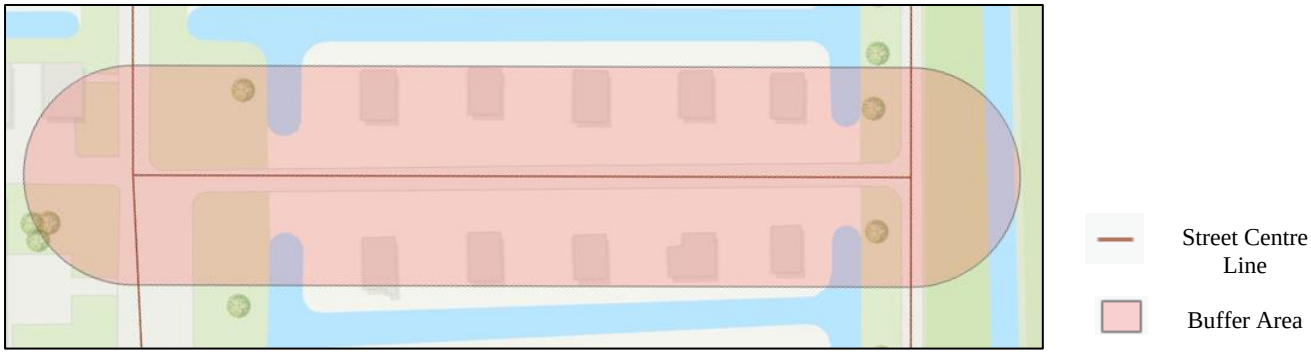


Figure 2. Buffer area used to capture urban features

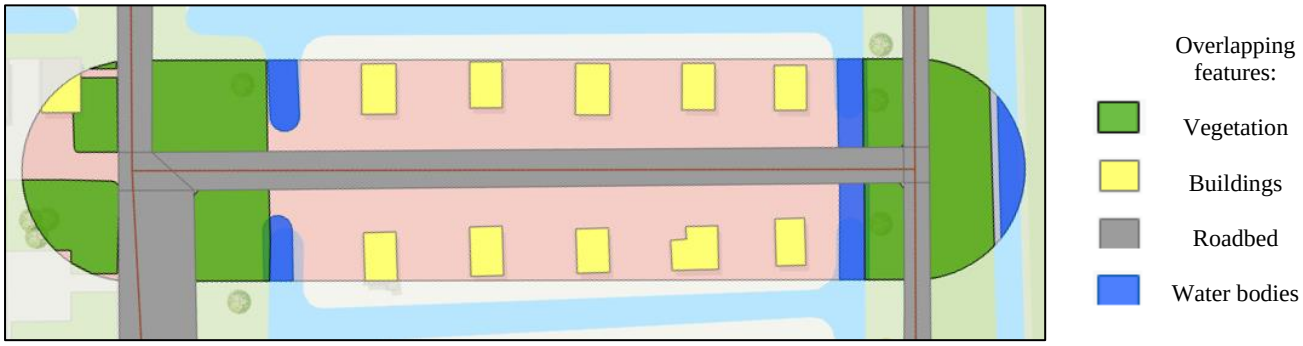


Figure 3. Isolation of urban features that are overlapping with the buffer area

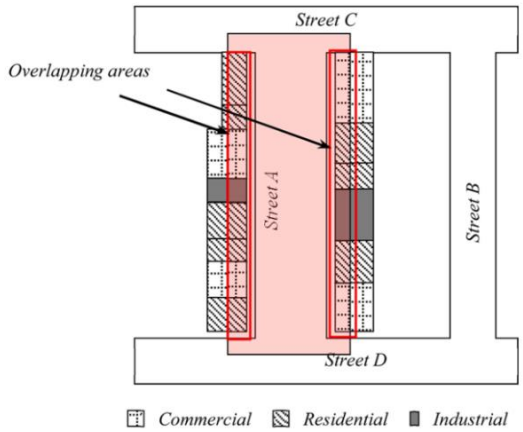


Figure 4. Determination dominant land use, adopted from (Pena Acosta et al., 2021a)

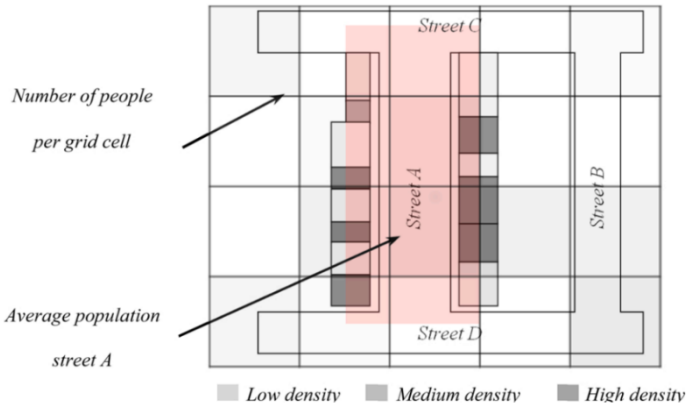


Figure 5. Determination population density, adopted from (Pena Acosta et al., 2021a)

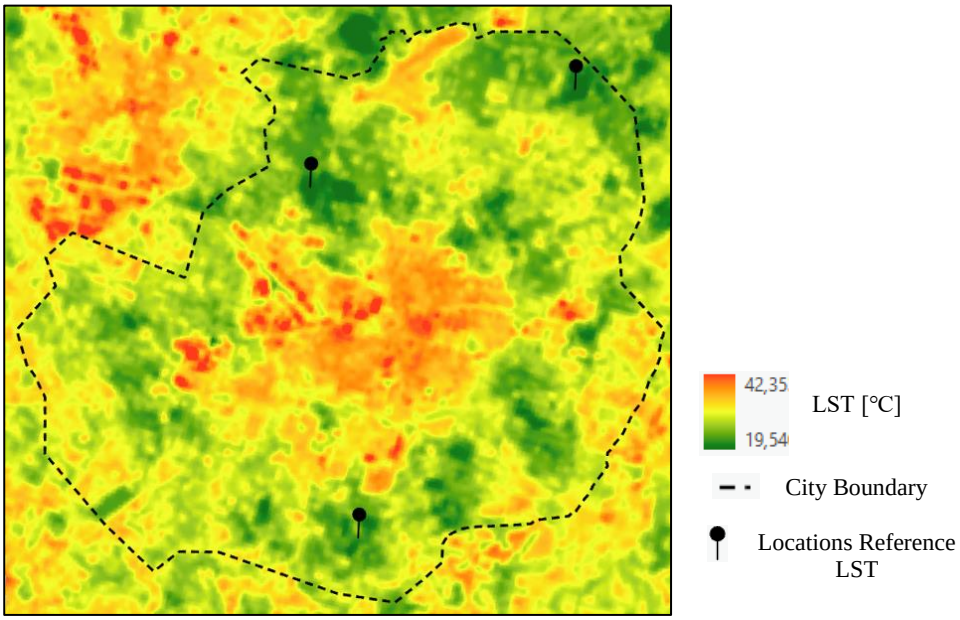


Figure 6. Heat map of the city of Enschede and locations of reference temperatures used to calculate ΔLST [%]

N = 7.143

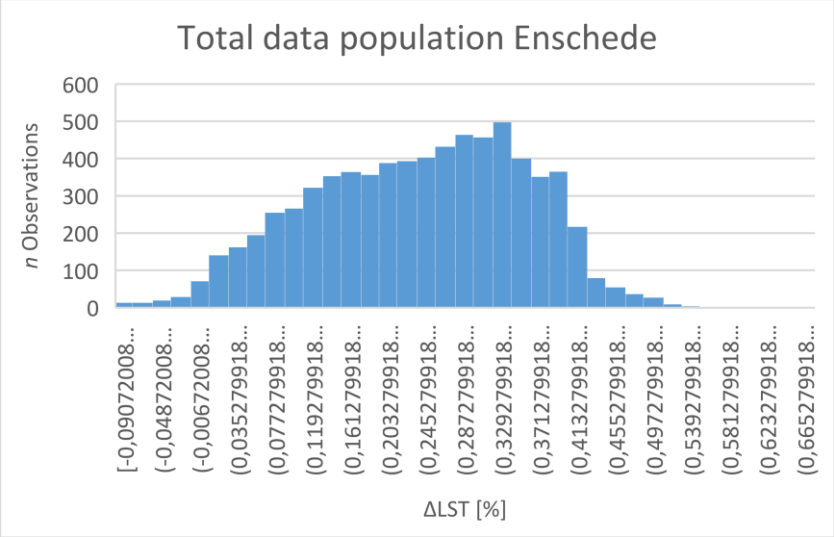


Figure 7. ΔLST [%] histogram, total data population Enschede

N = 6.692

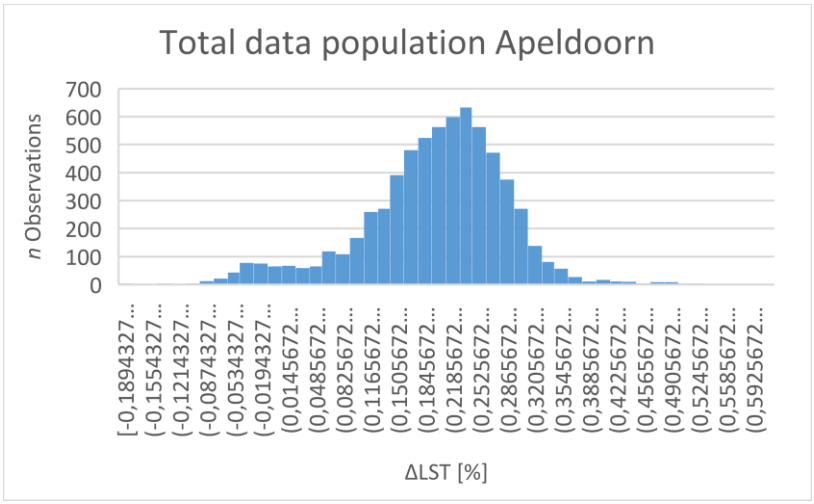


Figure 8. ΔLST [%] histogram, total data population Apeldoorn

N = 15.864

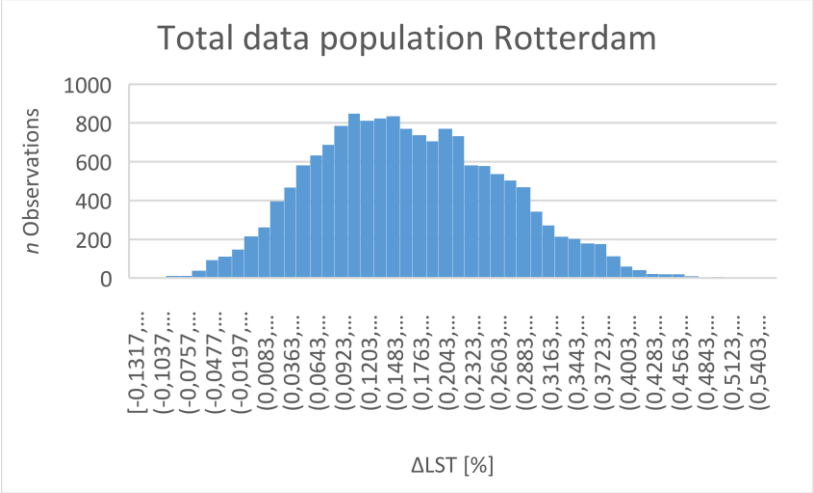
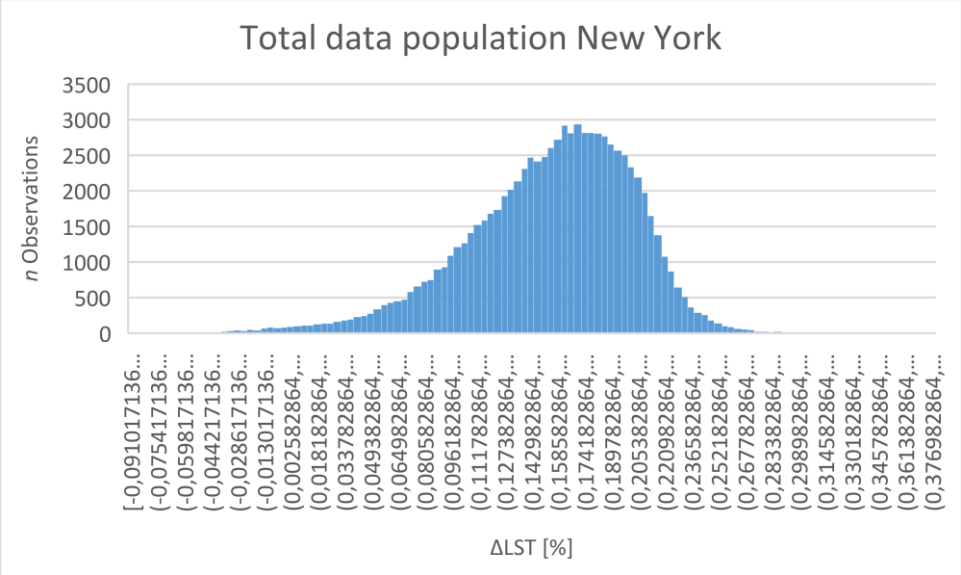
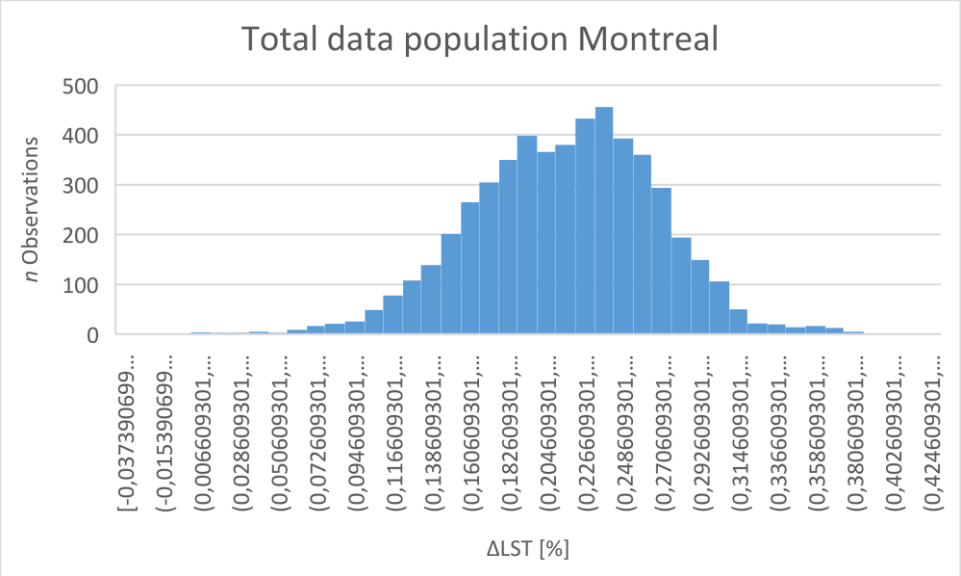


Figure 9. ΔLST [%] histogram, total data population Rotterdam



N = 83.822

Figure 10. ΔLST [%] histogram, total data population New York



N = 5.266

Figure 11. ΔLST [%] histogram, total data population Montreal

N = 4.763

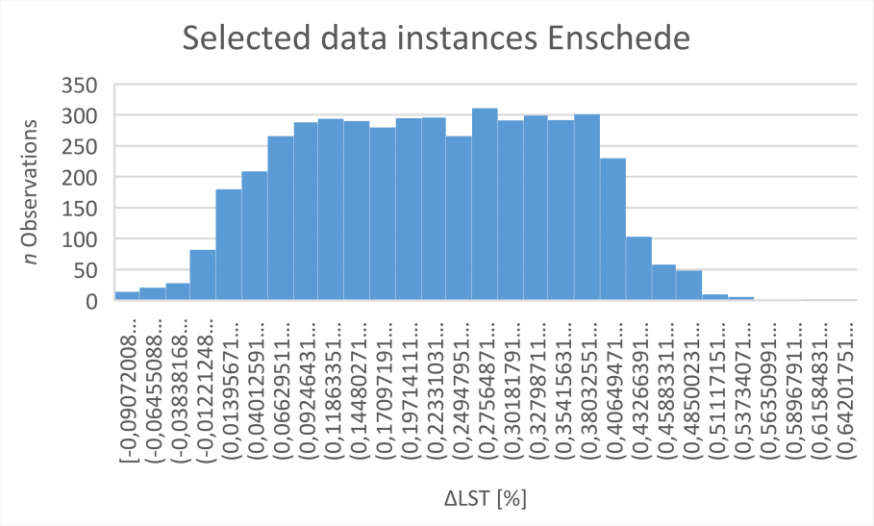


Figure 12. ΔLST [%] histogram, selected data population Enschede

N = 4.678

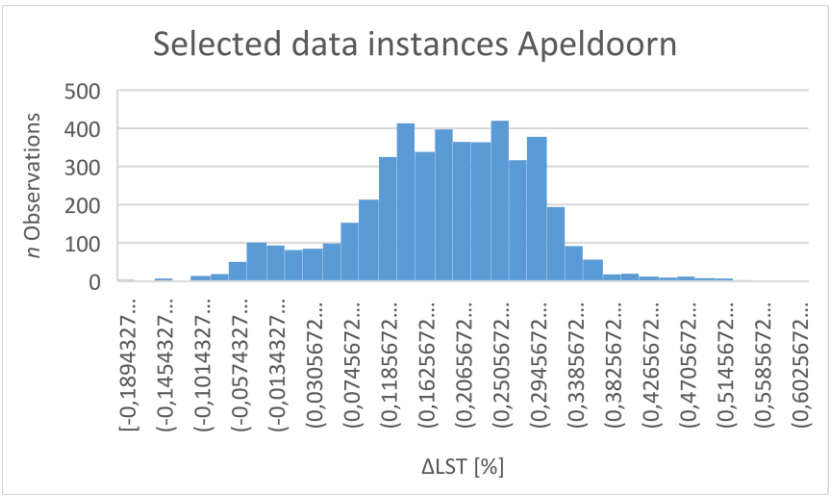


Figure 13. ΔLST [%] histogram, selected data population Apeldoorn

N = 4.640

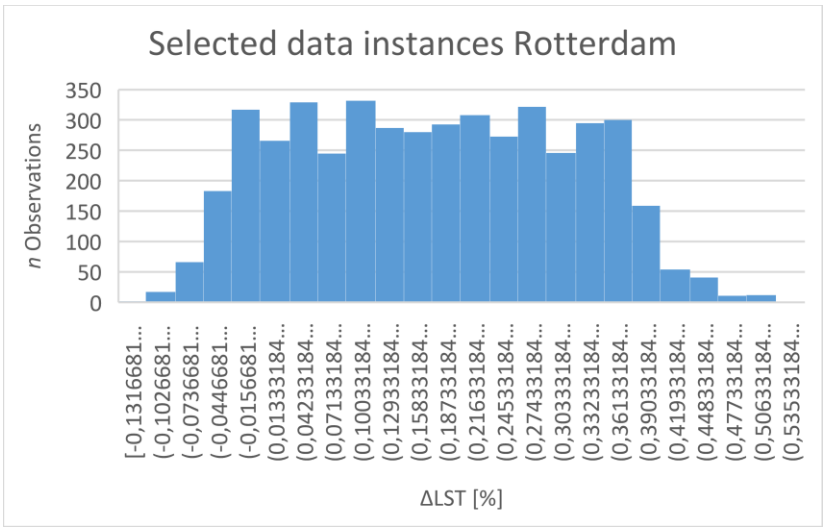
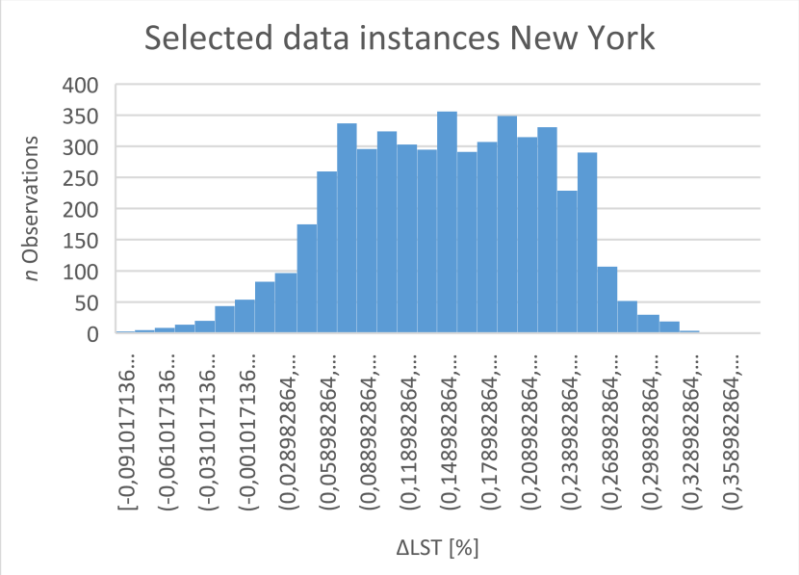
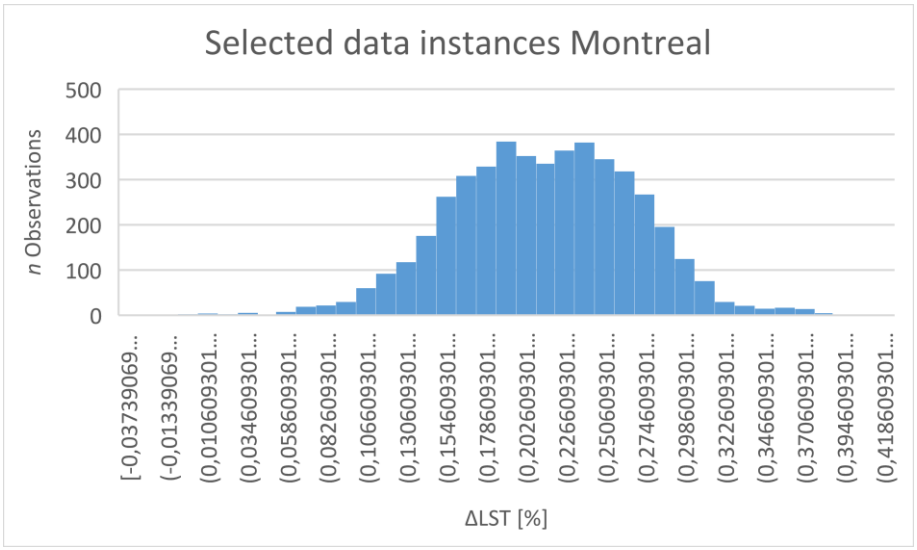


Figure 14. ΔLST [%] histogram, selected data population Rotterdam



N = 5.000

Figure 15. ΔLST [%] histogram, selected data population New York



N = 4.696

Figure 16. ΔLST [%] histogram, selected data population Montreal

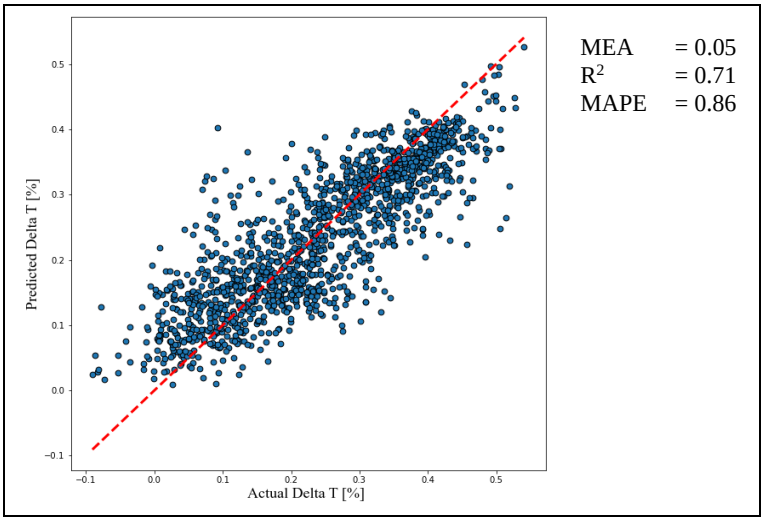


Figure 17(a). Scatter plot best performing RF, Enschede

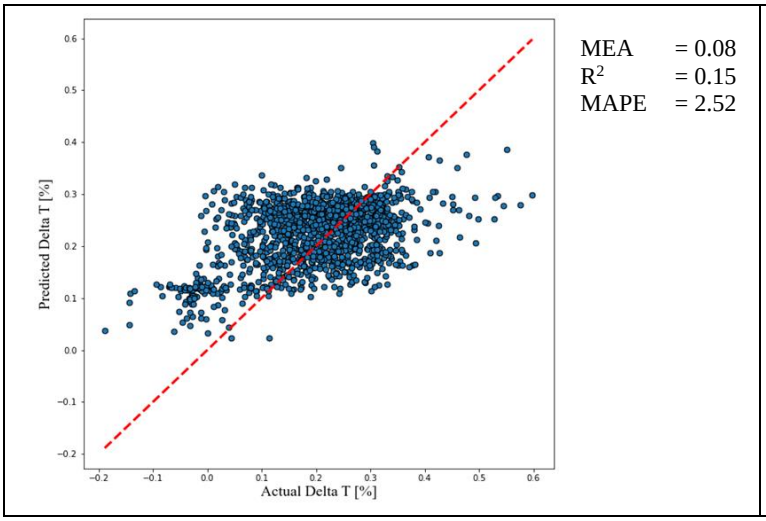


Figure 17(b). Model Enschede, tested on Apeldoorn

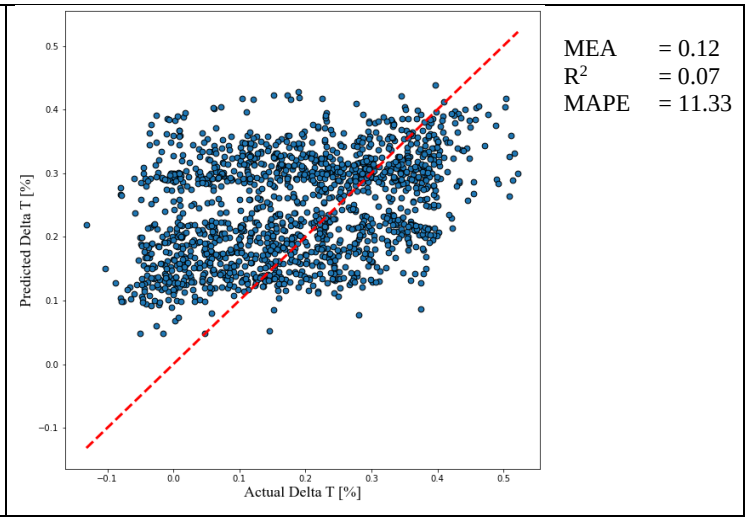


Figure 17(c). Model Enschede, tested on Rotterdam

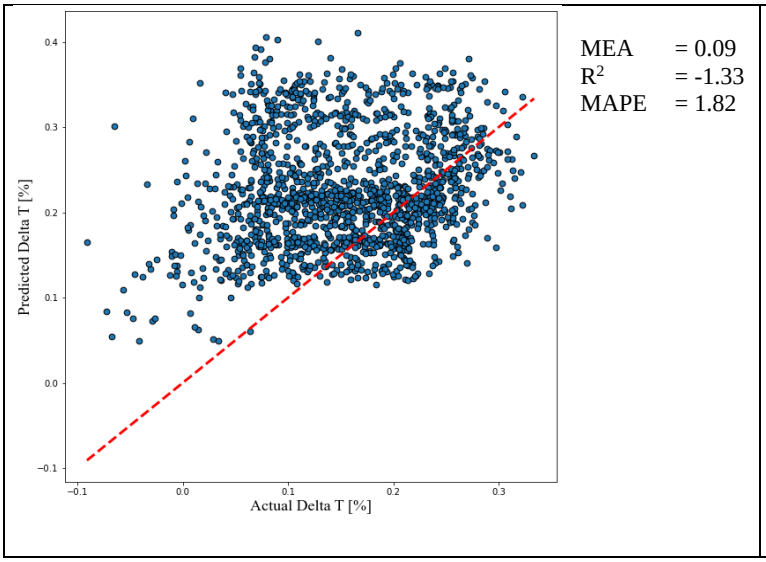


Figure 17(d). Model Enschede, tested on New York

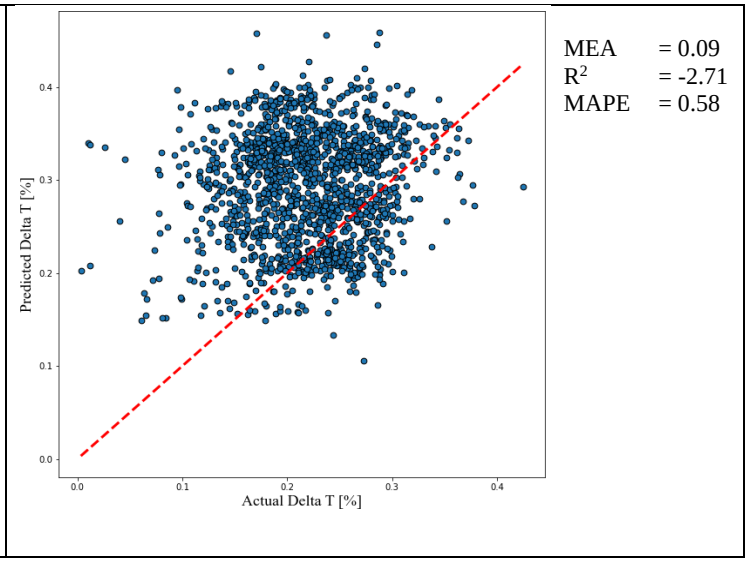


Figure 17(e). Model Enschede, tested on Montreal

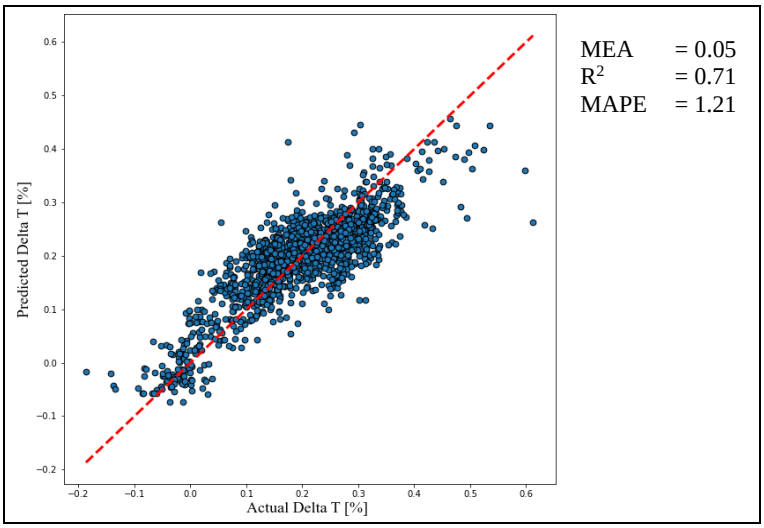


Figure 18(a). Scatter plot best performing RF, Apeldoorn

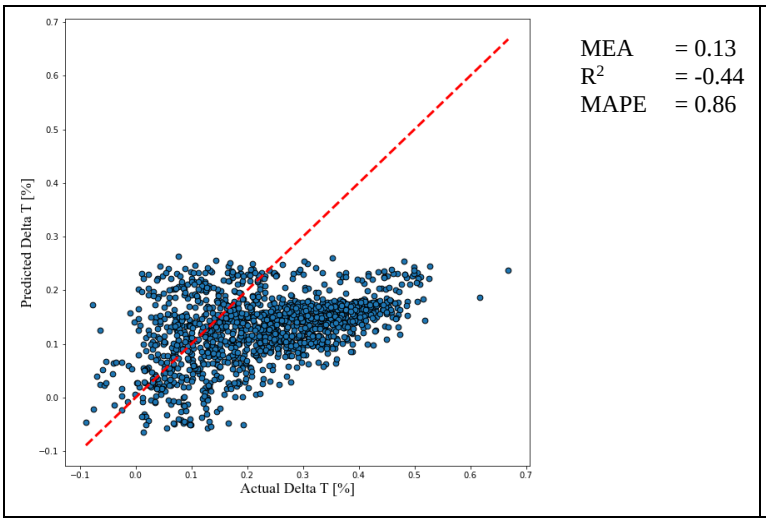


Figure 18(b). Model Apeldoorn, tested on Enschede

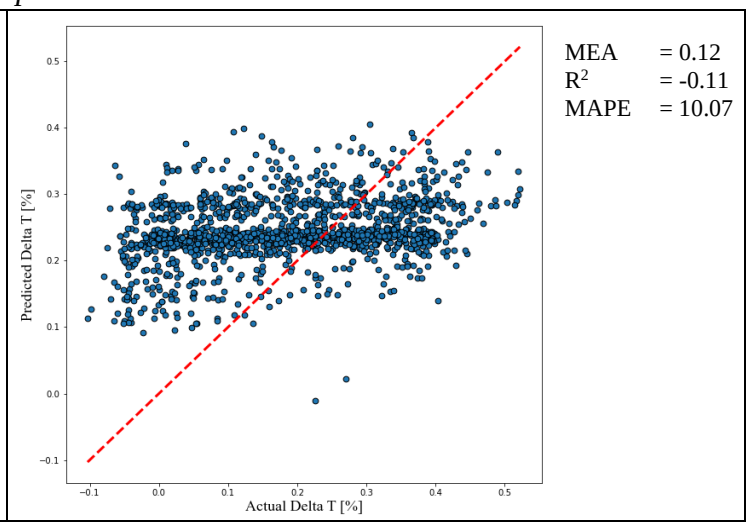


Figure 18(c). Model Apeldoorn, tested on Rotterdam

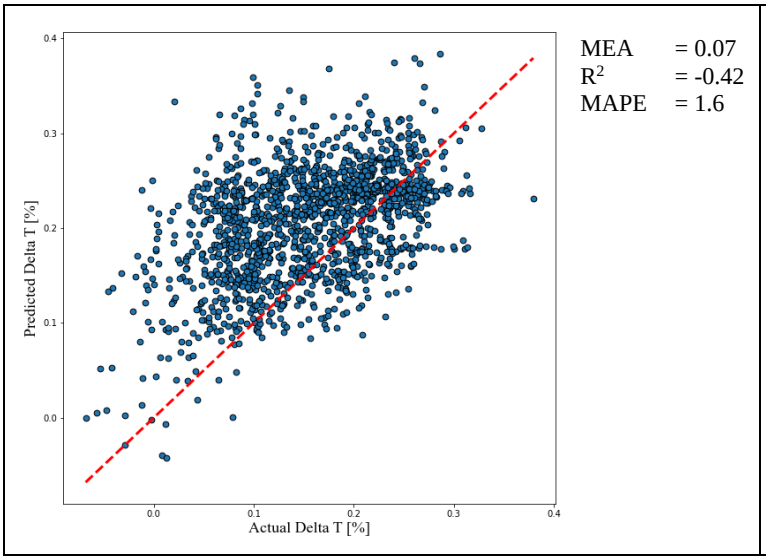


Figure 18(d). Model Apeldoorn, tested on New York

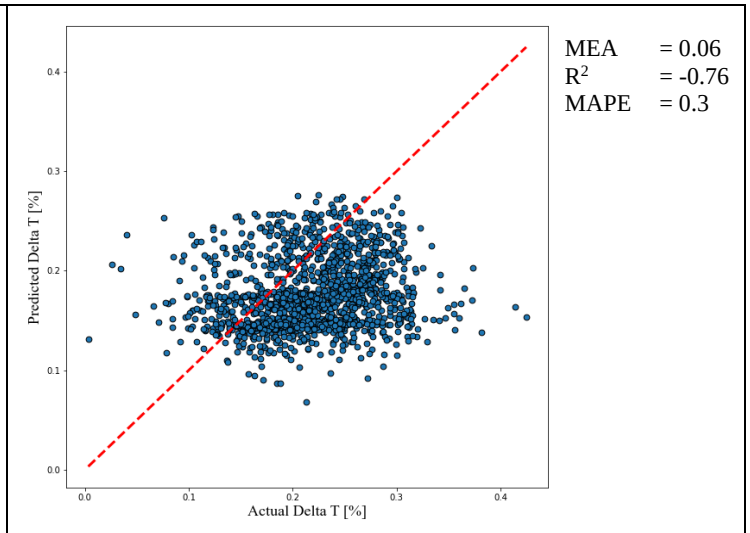


Figure 18(e). Model Apeldoorn, tested on Montreal

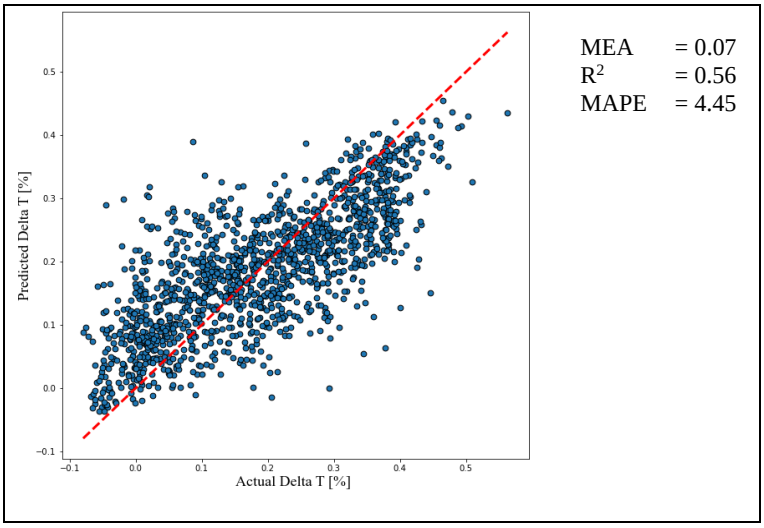


Figure 19(a). Scatter plot best performing RF, Rotterdam

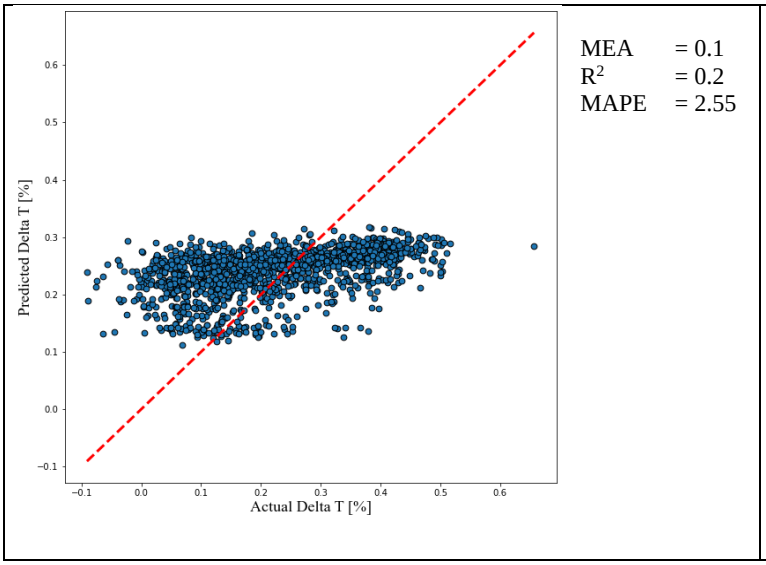


Figure 19(b). Model Rotterdam, tested on Enschede

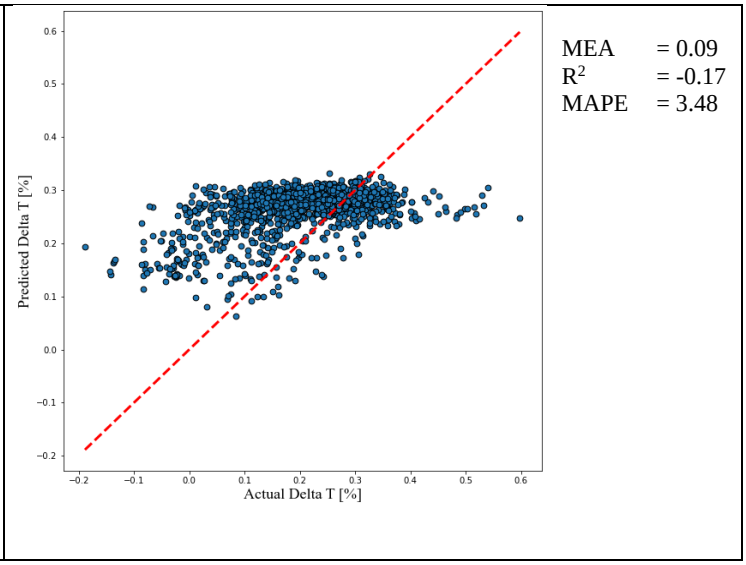


Figure 19(c). Model Rotterdam, tested on Apeldoorn

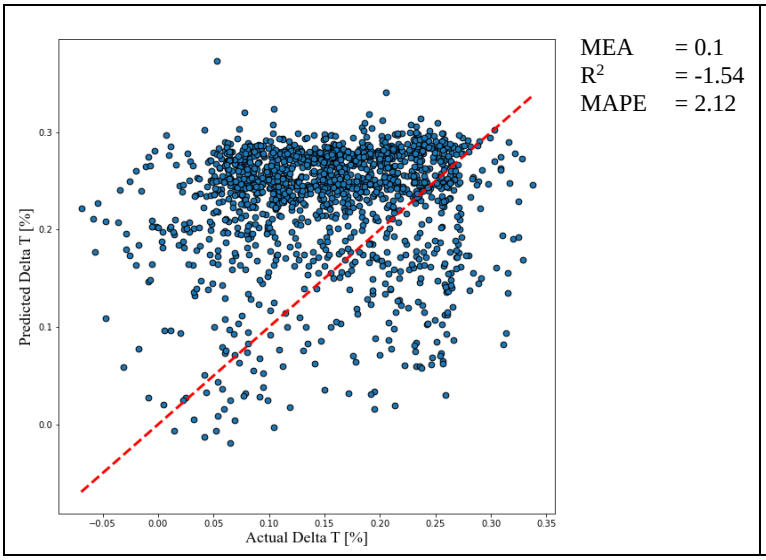


Figure 19(d). Model Rotterdam, tested on New York

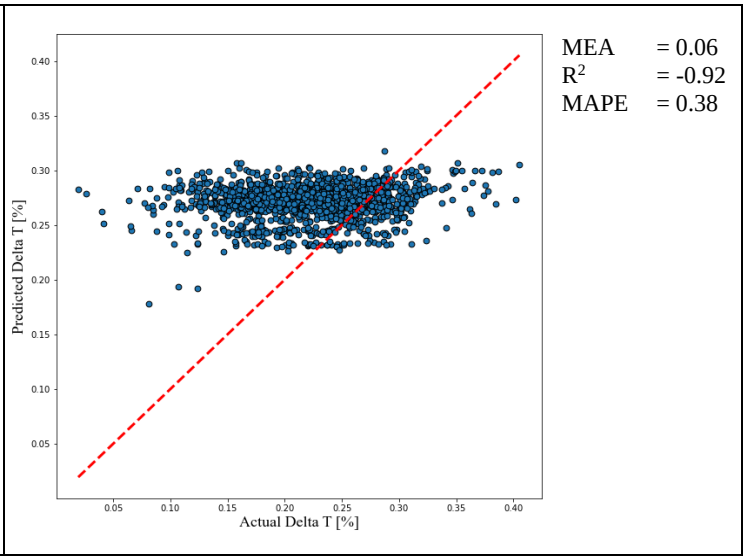


Figure 19(e). Model Rotterdam, tested on Montreal

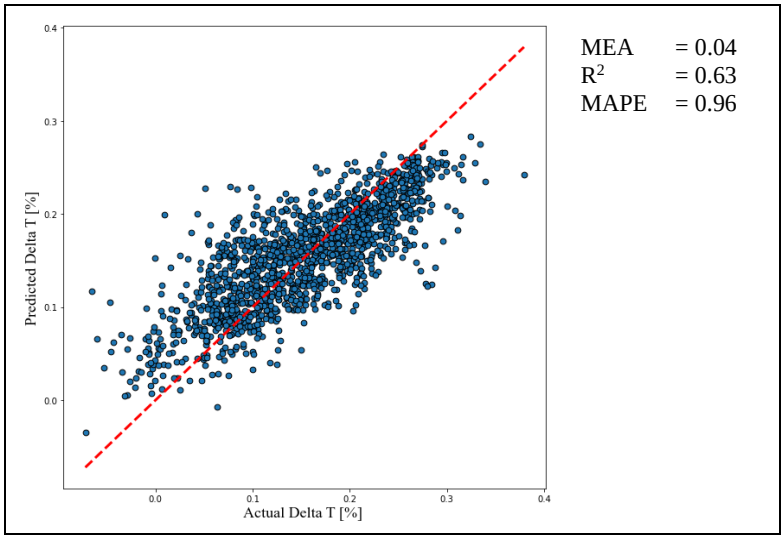


Figure 20(a). Scatter plot best performing RF, New York

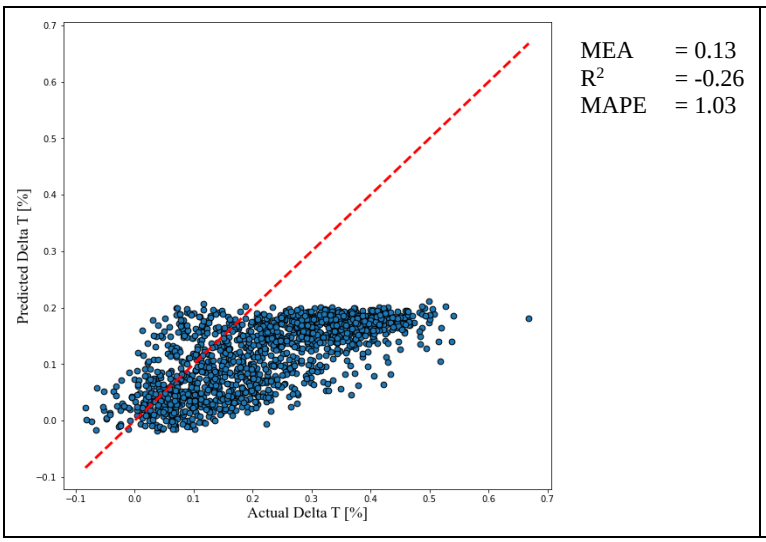


Figure 20(b). Model New York, tested on Enschede

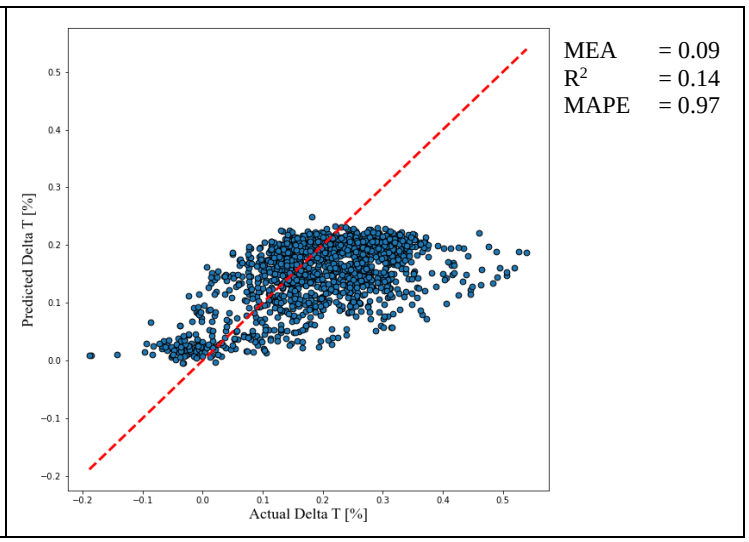


Figure 20(c). Model New York, tested on Apeldoorn

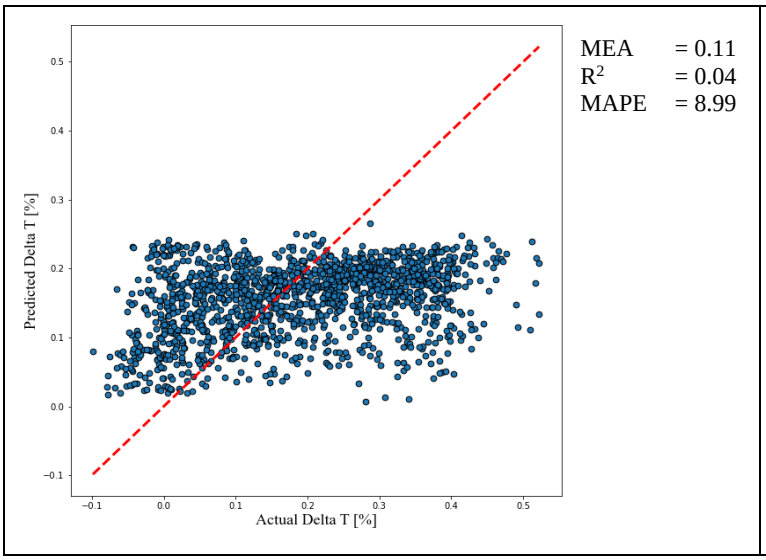


Figure 20(d). Model New York, tested on Rotterdam

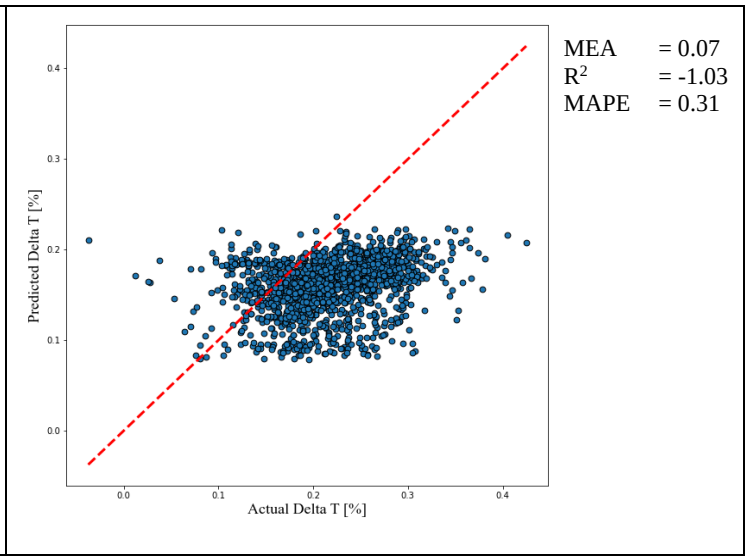


Figure 20(e). Model New York, tested on Montreal

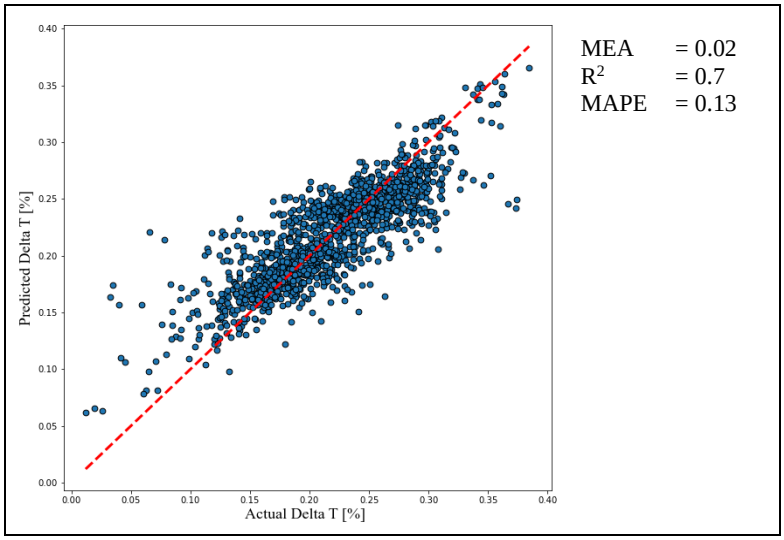


Figure 21(a). Scatter plot best performing RF, Montreal

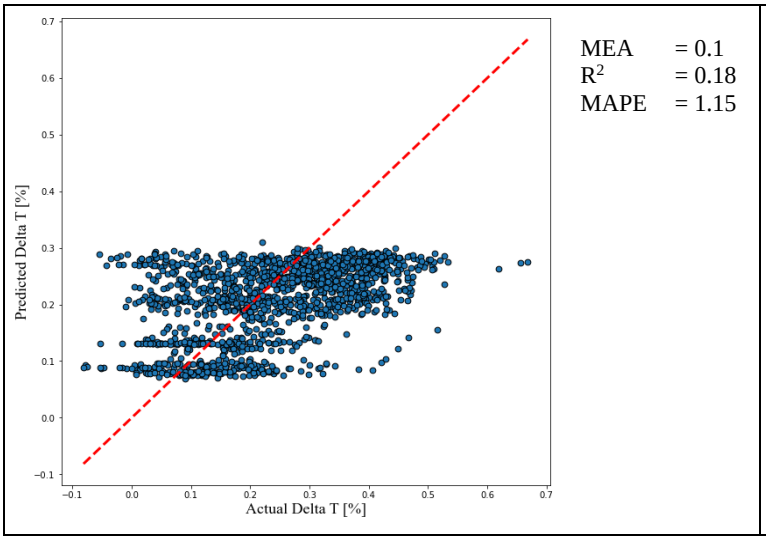


Figure 21(b). Model Montreal, tested on Enschede

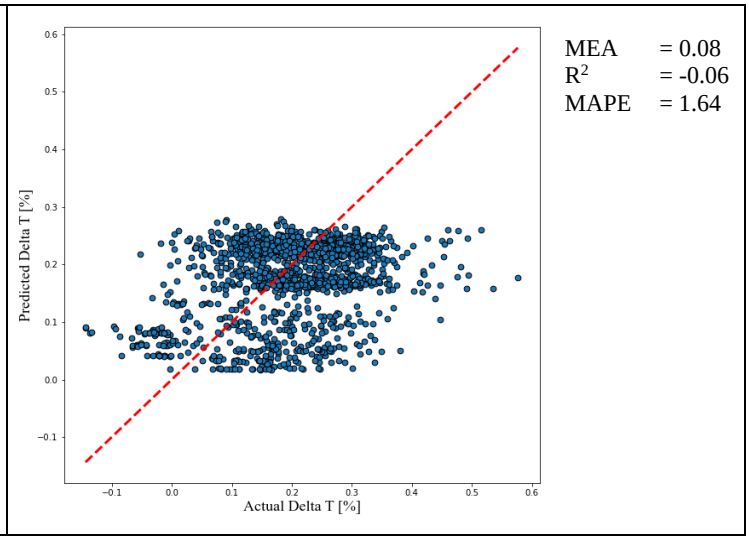


Figure 21(c). Model Montreal, tested on Apeldoorn

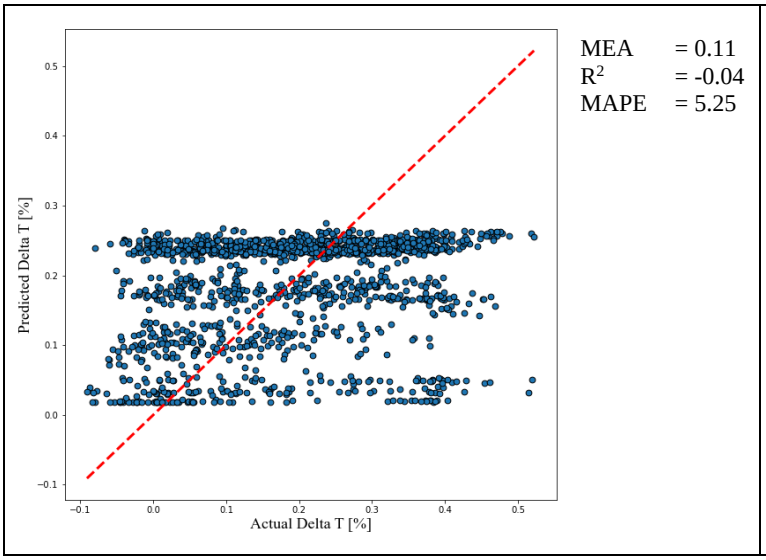


Figure 21(d). Model Montreal, tested on Rotterdam

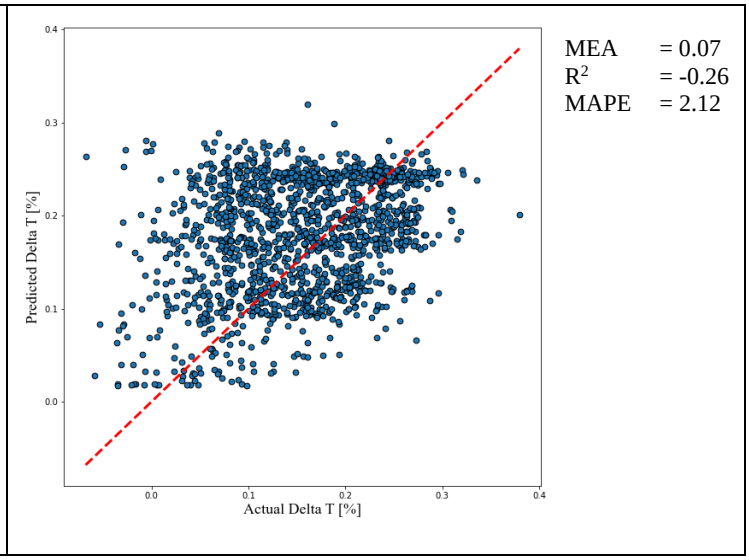
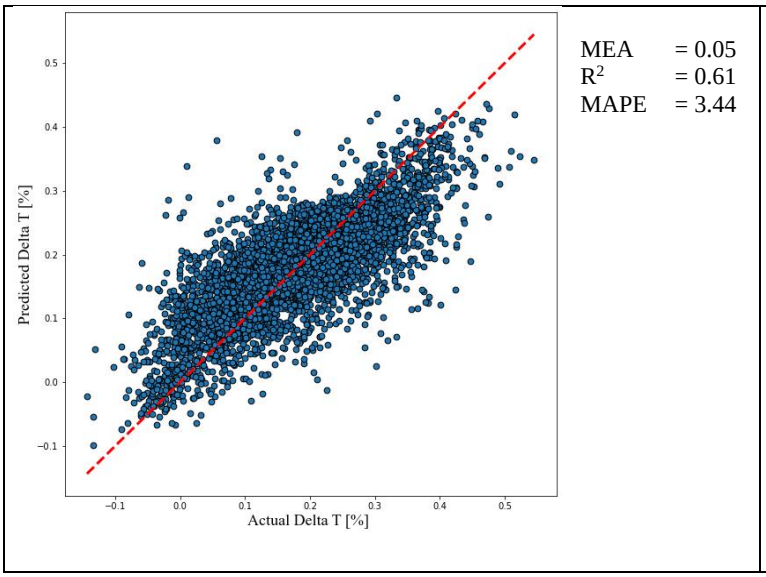
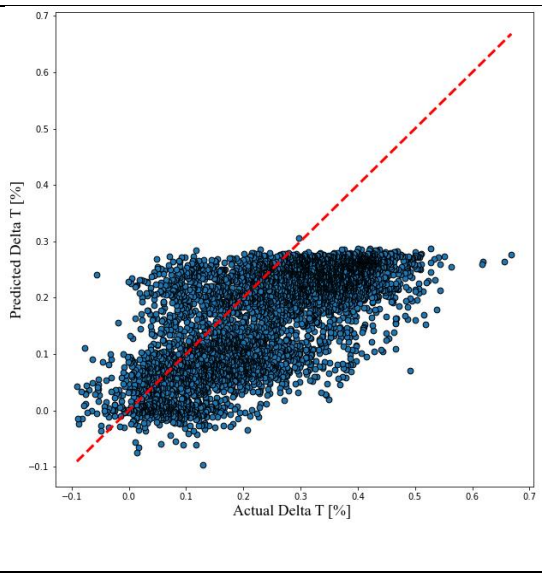


Figure 21(e). Model Montreal, tested on New York



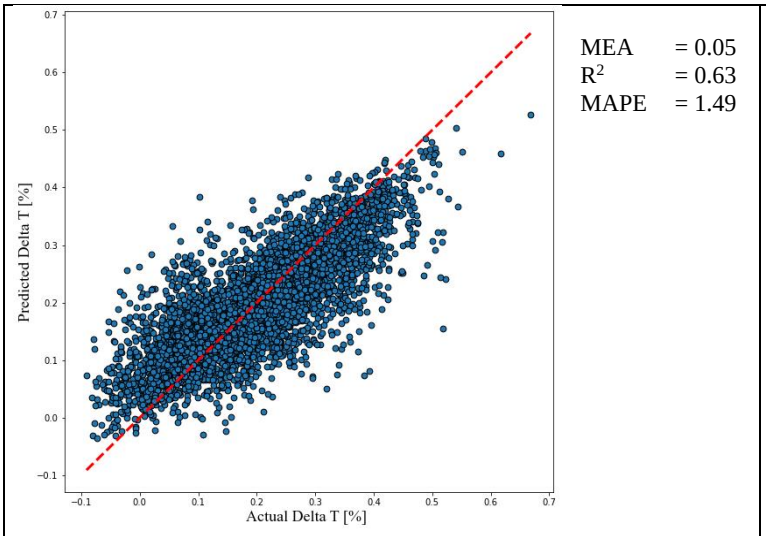
MEA = 0.05
R^2 = 0.61
MAPE = 3.44

Figure 22(a). Performance mixed dataset excl. Enschede



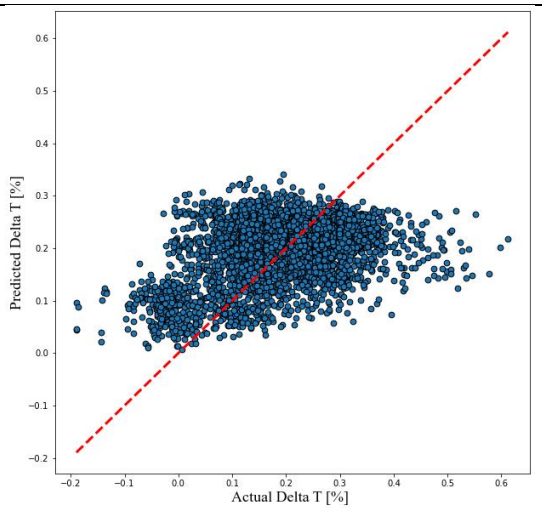
MEA = 0.1
R^2 = 0.14
MAPE = 1.1

Figure 22(b). Model tested on Enschede



MEA = 0.05
R^2 = 0.63
MAPE = 1.49

Figure 23(a). Performance mixed dataset excl. Apeldoorn



MEA = 0.08
R^2 = 0.15
MAPE = 2.24

Figure 23(b). Model tested on Apeldoorn

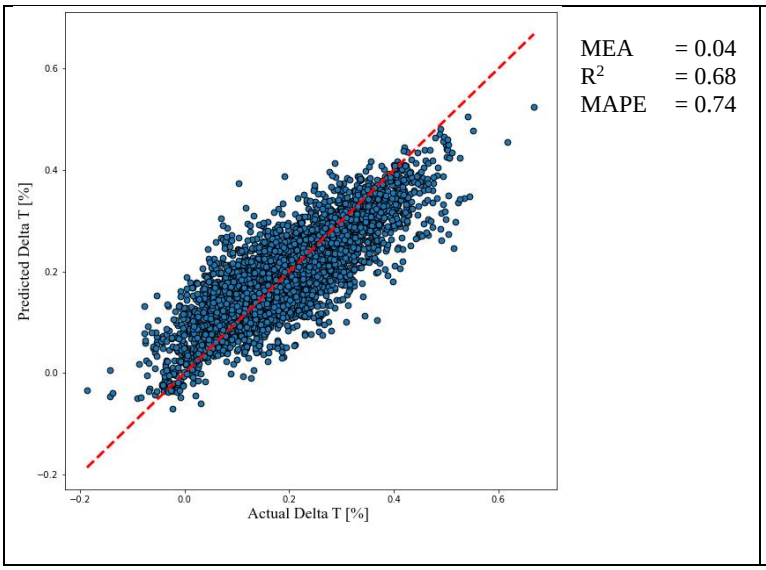


Figure 24(a). Performance mixed dataset excl. Rotterdam

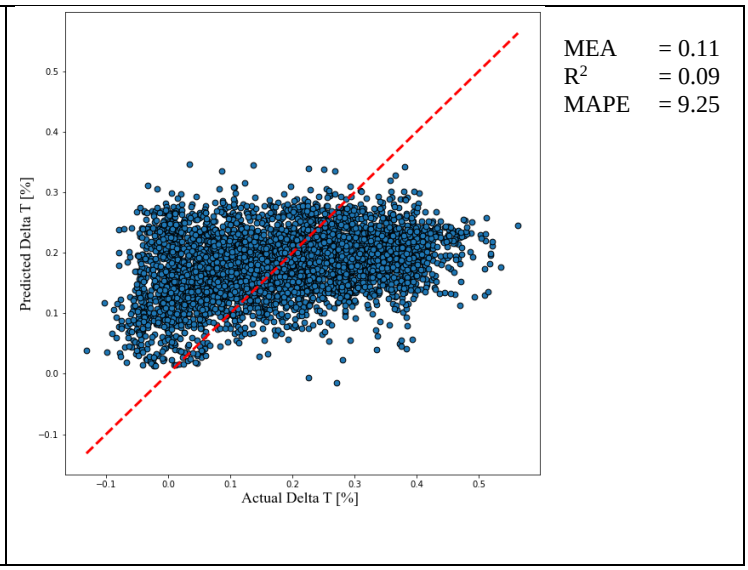


Figure 24(b). Model tested on Rotterdam

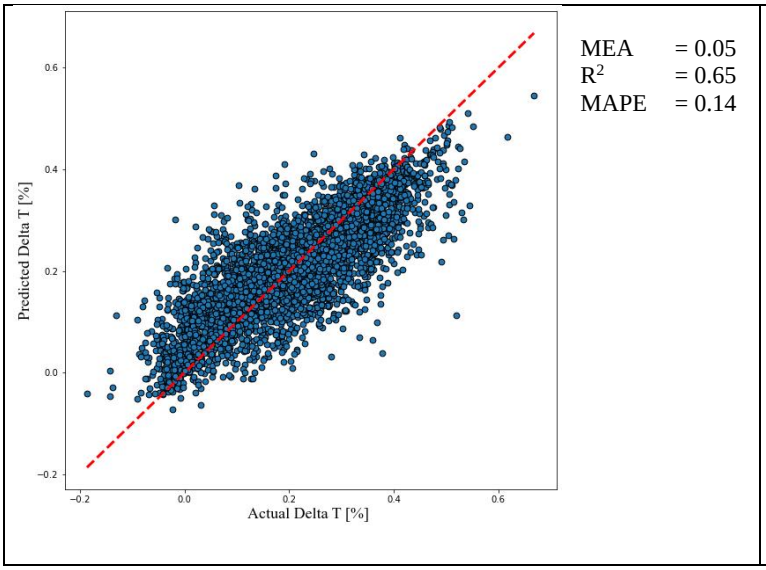


Figure 25(a). Performance mixed dataset excl. New York

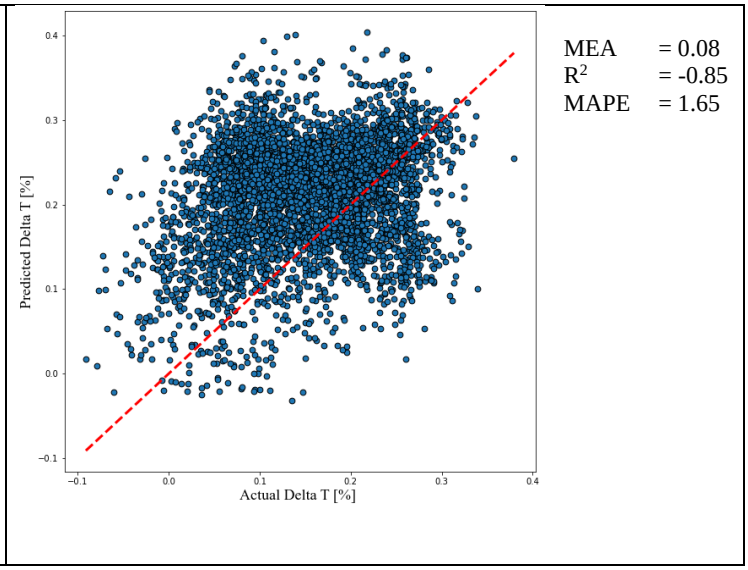


Figure 25(b). Model tested on New York

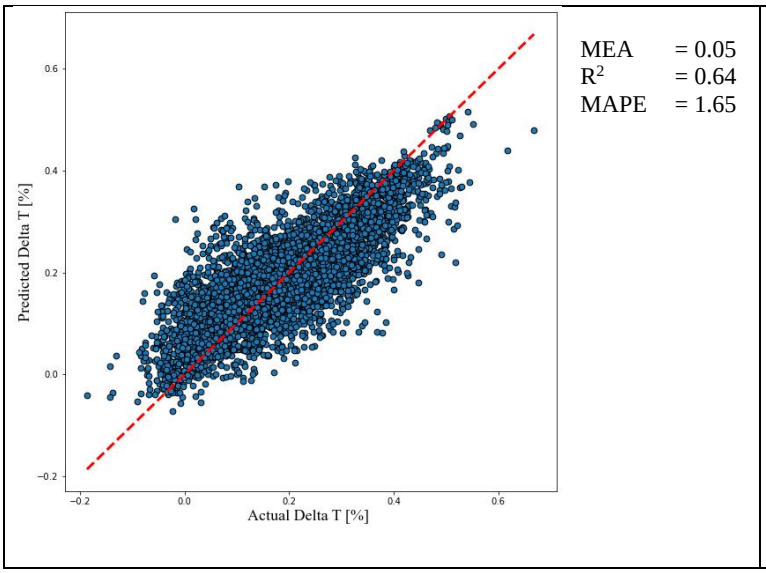


Figure 26(a). Performance mixed dataset excl. Montreal

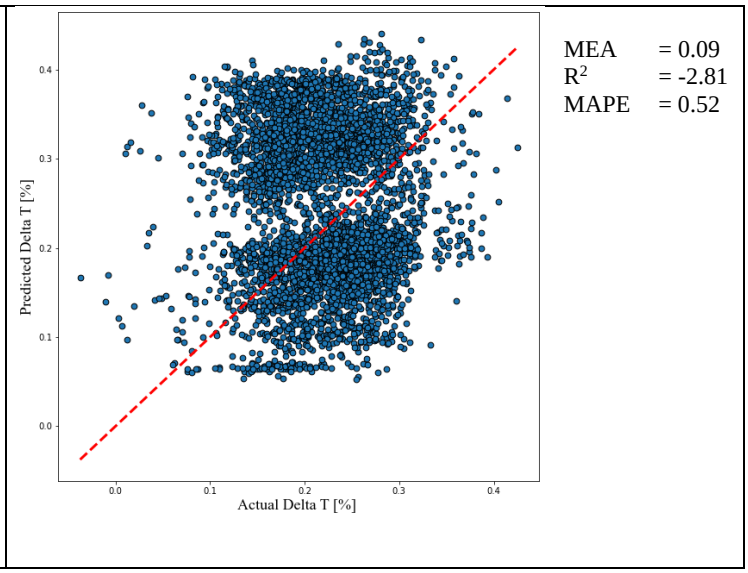


Figure 26(b). Model tested on Montreal

TABLES

Table 1. City selection for generalizability assessment

City	Regional climate	City area	Population
Enschede	Maritime	142 km^2	158.550
Apeldoorn		341 km^2	158.800
Rotterdam		324 km^2	623.652
New York	Humid Continental	784 km^2	8.419.000
Montreal		432 km^2	1.780.000

Table 2. Publicly available data sources to determine the feature values per city

Feature	Enschede	Apeldoorn	Rotterdam	New York	Montreal
Street Width	(3D Basisvoorziening - PDOK)			(CSCL | NYC Open Data)	(Montreal Open Data Portal)
Bld Density	(3D Basisvoorziening - PDOK)			(NYC 3D Model | NYC Open Data)	(Montreal Open Data Portal)
Veg Density	(3D Basisvoorziening - PDOK)			(DoITT)	(Montreal Open Data Portal)
Water Density	(3D Basisvoorziening - PDOK)			(DoITT)	(Montreal Open Data Portal)
Max Bld Height	(3D Basisvoorziening - PDOK)			(NYC 3D Model | NYC Open Data)	(Montreal Open Data Portal)
Mean Bld Height	(3D Basisvoorziening - PDOK)			(NYC 3D Model | NYC Open Data)	(Montreal Open Data Portal)
Std Bld Height	(3D Basisvoorziening - PDOK)			(NYC 3D Model | NYC Open Data)	(Montreal Open Data Portal)
Elevation	(3D Basisvoorziening - PDOK)			(DEM | NYC Open Data)	(Montreal Open Data Portal)
HW Ratio	(3D Basisvoorziening - PDOK)			(CSCL | NYC Open Data)	(Montreal Open Data Portal)
Mean Population	(Centraal Bureau Statistiek)			(WorldPop - Population Counts)	(WorldPop - Population Counts)
Predominant Land Use	(Geo Services - PDOK)			(PLUTO - NYC DCP)	(Montreal Open Data Portal)

Table 3. Land use categorization

Land Use Category	Label
Parks and conversation	0
Industrial	1
Residential	2
Commercial	3
Governmental	4
Agricultural	5
Paved open spaces	6
Mix	7

Table 4. Example of structured dataset, including all features and dependent variable ΔLST [%]

Street Width	Bld Density	Veg Density	Water Density	Max Bld Height	Mean Bld Height	Std Bld Height	Elevation	HW Ratio	Mean Population	Predominant Land Use		ΔLST [%]
7,62	0,13	0,26	0,00	10,40	8,98	2,00	3,21	1,17	0,00	7		0,054
6,09	0,02	0,80	0,00	105,91	104,14	1,82	103,28	17,08	7,61	2		0,044
5,48	0,00	0,68	0,15	0,00	0,00	0,00	0,71	0,00	19,73	0		0,054
…	…	…	…	…	…	…	…	…	…	…		…

Table 5. Example calculation Similarity Indeces for Enschede

Enschede		Street Width	Bld Density	Veg Density	Water Density	Max Bld Height	Mean Bld Height	Std Bld Height	Elevation	HW Ratio	Mean Population	Predominant Land Use	**Similarity Index**
Feature Importance (FI)		0,04	0,07	0,33	0,02	0,06	0,02	0,02	0,12	0,02	0,20	0,11	
Apeldoorn	KS-Statistic	0,363	0,203	0,204	0,103	0,469	0,400	0,171	0,842	0,172	0,414	0,210	
	1 – KS	0,637	0,797	0,796	0,897	0,531	0,600	0,829	0,158	0,828	0,586	0,790	
	FI * (1 – KS)	0,025	0,056	0,263	0,018	0,032	0,012	0,017	0,019	0,017	0,117	0,087	**0.66**
Rotterdam	KS-Statistic	0,223	0,071	0,268	0,061	0,222	0,223	0,099	1,000	0,077	0,240	0,195	
	1 – KS	0,777	0,929	0,732	0,939	0,778	0,777	0,901	0,000	0,923	0,760	0,805	
	FI * (1 – KS)	0,031	0,065	0,242	0,019	0,047	0,016	0,018	0,000	0,018	0,152	0,089	**0.70**
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Table 6. Total data populations per city

City	**Data Population**
Enschede	7.143
Apeldoorn	6.692
Rotterdam	15.864
New York	83.822
Montreal	5.266

Table 7. Feature importances, derived from best perfoming RFs per city

Feature Importance

	Street Width	Bld Density	Veg Density	Water Density	Max Bld Height	Mean Bld Height	Std Bld Height	Elevation	HW Ratio	Mean Population	Predominant Land Use
Enschede	0,04	0,07	0,33	0,02	0,06	0,02	0,02	0,12	0,02	0,2	0,11
Apeldoorn	0,07	0,06	0,27	0,01	0,03	0,02	0,03	0,3	0,03	0,03	0,16
Rotterdam	0,07	0,09	0,11	0,04	0,08	0,04	0,04	0,3	0,03	0,1	0,12
New York	0,06	0,07	0,22	0,03	0,1	0,09	0,05	0,11	0,07	0,13	0,06
Montreal	0,05	0,04	0,17	0	0,04	0,06	0,03	0,37	0,04	0,08	0,1
Average	0,063	0,065	0,193	0,02	0,063	0,053	0,038	0,27	0,043	0,085	0,11
Mix	0,06	0,04	0,24	0,02	0,05	0,05	0,03	0,19	0,03	0,12	0,08

Table 8. Similarity Indeces

		Testing				
	Similarity Index	Enschede	Apeldoorn	Rotterdam	New York	Montreal
Training	Enschede	**1,00**	0,66	0,70	0,71	0,62
	Apeldoorn	0,59	**1,00**	0,54	0,73	0,59
	Rotterdam	0,59	0,52	**1,00**	0,62	0,47
	New York	0,63	0,65	0,66	**1,00**	0,47
	Montreal	0,60	0,48	0,41	0,58	**1,00**

Table 9. Test results of external cross-validation of best performing RFs

		Testing				
	MAE R-squared MAPE	Enschede	Apeldoorn	Rotterdam	New York	Montreal
Training	Enschede	**0,05** **0,71** **0,86**	0,08 0,15 2,52	0,12 0,07 11,33	0,09 -1,33 1,82	0,09 -2,71 0,58
	Apeldoorn	0,13 -0,44 0,86	**0,05** **0,71** **1,21**	0,12 -0,11 10,07	0,07 -0,42 1,6	0,06 -0,76 0,3
	Rotterdam	0,1 0,2 2,55	0,09 -0,17 3,48	**0,07** **0,56** **4,45**	0,1 -1,54 2,12	0,06 -0,92 0,38
	New York	0,13 -0,26 1,03	0,08 0,14 0,97	0,11 0,04 8,99	**0,04** **0,63** **0,96**	0,07 -1,03 0,31
	Montreal	0,1 0,18 1,15	0,08 -0,06 1,64	0,11 -0,04 5,25	0,07 -0,26 2,12	**0,02** **0,7** **0,13**

Table 10. Correlations between Similarity Index and performance metrics

	MAE	**R^2**	**MAPE**
Coefficient (r)	0,011	-0,144	0,024
N	20	20	20
T statistic	0,049	-0,616	0,104
DF	18	18	18
p value	0,960	0,544	0,919