


Ensemble Learning Traffic Model for Sofia: A Case Study

Danail Brezov ^{1,*},† and Angel Burov ^{2,†} ¹ Department of Mathematics, University of Architecture, Civil Engineering and Geodesy, 1164 Sofia, Bulgaria² Department of Urban Planning, University of Architecture, Civil Engineering and Geodesy, 1164 Sofia, Bulgaria

* Correspondence: danail.brezov@gmail.com

† Current address: 1 Hristo Smirnenski Blvd., University of Architecture, Civil Engineering and Geodesy, 1164 Sofia, Bulgaria.

Featured Application: Optimizing urban traffic/logistics; estimating pollution due to combustion engines; studying the relation between urban pollution and respiratory/cardiovascular problems.

Abstract: Traffic models have gained much popularity in recent years, in the context of smart cities and urban planning, as well as environmental and health research. With the development of Machine Learning (ML) and Artificial Intelligence (AI) some limitations imposed by the traditional analytical, numerical and statistical methods have been overcome. The present paper shows a case study of traffic modeling with scarce reliable data. The approach we propose resorts on the advantages of ensemble learning using a large number of related features such as road and street categories, population density, functional analysis, space syntax, previous traffic measurements and models, etc. We use advanced regression models such as Random Forest, XGBoost, CatBoost etc., ranked according to the chosen evaluation metrics and stacked in a weighted ensemble for optimal fitting. After a series of consecutive data imputations we estimate the annual average daily traffic distribution in the street and road network of Sofia city and the metropolitan municipality for 2018 and 2022, and the NO₂ levels for 2021 with accuracy resp. 78%, 74% and 92%, using AutoGluon and Scikit-Learn.

Keywords: urban traffic models; machine learning; multiple regression; data imputation; AutoML**Citation:** Brezov, D.; Burov, A.Ensemble Learning Traffic Model for Sofia: A Case Study. *Appl. Sci.* **2023**, *13*, 4678. <https://doi.org/10.3390/app13084678>

Academic Editor: Vincent A. Cicirello

Received: 1 March 2023

Revised: 23 March 2023

Accepted: 30 March 2023

Published: 7 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Transport models for urban and intercity traffic are a valuable tool for optimization and planning of mobility and accessibility, estimation of environmental pressures, sources and dispersion of chemical pollutants, such as carbon monoxide, sulfur dioxide and nitrogen oxides, inhalable particulate matter, as well as more precise evaluation of traveling time. Machine learning and deep learning algorithms provide increasingly accurate results in this field (we refer to [1,2] for a brief review on the subject). In particular, ensemble learning algorithms [3] prove quite useful in traffic, pollution or landscape models with lots of fuzzy data. They are less prone to over-fitting than neural networks, do not require extensive preprocessing, allow for nice hyper-parameter tuning and seem to be immuned against problems like multicollinearity. Our most preferred tools are Random Forest (RF) which has become the standard in this field [4], Extreme Gradient Boosting (XGB) that became famous in Kaggle competitions shortly after its release, as well as the WeightedEnsembleL2 algorithm proposed by the Python AutoML module (AutoGluon) we use for ranking, optimization and training of the models. AutoML tools are intuitive to work with and rather fast, but more importantly they provide a convenient framework for us to conduct more complex studies, e.g., considering the multi-modal setting [5] in which satellite or drone images are used along with tabular continuous, categorical and text data. Such models would have been quite difficult to implement with only the standard data science libraries, but lately they have become a standard approach in the field [6]. This study, however, relies on the classical tabular data of various sorts, as a case study, in the spirit

of [7,8]. Namely, we work with estimates of average traffic values, focusing only on spatial, not temporal distribution and use multi-feature regression to impute the missing values. Some studies rely on probabilistic methods, such as PCA [9,10], deep learning [11] or Markov chains [12]. This article evaluates the performance of ‘bagging’ (RF), ‘boosting’ (XGB) and ‘stacking’ (weighted ensemble) algorithms which have been studied for the purposes of both applied and fundamental research tasks related to air pollution, health and urban development performed in the context of intermediate scarcity of data, hardware, software, human and financial resources for the related degrees of complexity. There are few studies that use the ensemble learning approach for traffic related issues, more often oriented towards traffic speed [13], congestion [14] and traffic flow prediction [15], while in [7] the authors use a similar approach for a country-scale annual average daily traffic estimate. Thus, we believe our study fills a gap in the contemporary urban traffic ML modeling, especially related to Eastern European cities and other similar case studies.

The case of Sofia in the context of Bulgaria and South-Eastern Europe can be described as exemplary for the low level from which the city started its environmental, green urban transition [16] and the many but still rarely well coordinated political and expert-led efforts resulting to moderate or even seemingly strong achievements [17] on the one hand, while on the other, the persistence of traffic air pollution and related problems, as well as low availability of traffic data for actual transport modeling. This includes discontinuities and many inconsistencies in data gathering, storing and management in a semi-transparent and relatively poorly communicated process among various institutions, experts and the scientific community. Our sincere hope is to see some significant improvement in the coming years, so that the ML algorithms could become even more helpful in these digitally underdeveloped cities for purposes of traffic optimization, environmental and especially air pollution screening.

The article begins with a brief preliminary section explaining the basic principles behind ensemble learning algorithms, followed by description of the materials and methods used, including the data sources, pre-processing and design of the ML algorithms. Their performance and the quality of the results is commented in a separate section and we end with a discussion attempting to position our case study in the broader context of traffic modelling and related issues such as pollution and public health in large urban areas.

2. Preliminaries: Ensemble Learning Algorithms

In this section we provide a brief intuitive explanation of the main concepts in the ensemble learning techniques we rely on for tabular prediction in our study. However, they are quite promising also as classification algorithms and perform nicely in the multimodal setting combining text, tabular and image/sound data (see [5] for more details).

2.1. Decision Trees and Random Forests

ML and AI algorithms perceive the world as a black box model—pretty much as humans do, at least beyond the scale of their everyday experience and scientific knowledge. Both humans and machines tend to learn and improve from their mistakes or wrong predictions—the efficiency of this process (not the absence of errors) is how we usually measure intelligence, be it natural or artificial. This corrective process has different mechanisms to manifest itself—from dopamine cycles in mammals to ‘back-propagation’ in neural networks. Both units, however, have the tendency to be either too broad (underfitting) or too specific for the training data (overfitting) in their estimates when facing a problem with overwhelming complexity, e.g., making a reliable weather forecast. This phenomenon is known as the ‘bias-variance trade-off’. One possible strategy to make the model more inclusive without sacrificing much precision is to allow the major decisions to be taken by some sort of democratic voting process. Typically the atomic constituents of an ensemble model would be decision trees—relatively simple algorithms using straightforward binary logic derived from experience to predict outcomes in unknown situations, e.g., a simple (multi-label) classification tree would rate a person ‘overweight’ if their body mass index (BMI) exceeds 25 and ‘underweight’ if is below 18. For more complex

decisions, however, single trees tend to overfit drawing absurd conclusions like ‘the last time we did not please the gods, there was draught, so we need human sacrifice’. Ensemble learning is a democratic concept that embraces the imperfection of individual trees and turns it into an advantage—a large group of ‘weak learners’ has better chances of success than a single ‘strong learner’, pretty much like in wild nature. But if all trees are fed the same data and obey identical set of binary rules, they would simply behave as clones, not contributing to the collective advancement. So, weak learners are trained on different datasets, obtained from the original one via bootstrapping (random sampling with replacement), so they have different backgrounds. To ensure even more diversity of opinion, ensemble learning also gives different ‘way of thinking’ for individual trees by providing them with different subsets of features. In the end, certain trees turn out to be better adapted for certain situations, some turn out to be rather ‘brilliant’, others—quite ‘dumb’, but this way the collective decision is well protected against both underfitting and overfitting—this whole process is known as ‘bagging’. In the Random Forest algorithm for instance, it is taken by majority vote in the case of classification and simple averaging for regression problems. This seemingly primitive Greek-type democracy provides great results in rather complex problems involving traffic, weather, ecology, social phenomena etc. Speed is a major issue in this setting, but the algorithm is perfectly suited for parallel processing, so it scales nicely.

2.2. Bagging vs. Boosting vs. Stacking

An alternative to the horizontal equity-type structure of Random Forest is relying on natural hierarchies to advance the system. The idea is to start with a simple base model trained on all available features and after accessing its performance, run it again but with biased data, shifted towards instances, for which it failed to give good predictions. This process is repeated over and over, like a karma cycle, until it is sufficiently refined, and the final result is given as a weighted sum of the predictions of all its versions (reincarnations). This process is known as ‘boosting’ and typically works with decision trees with specific structure in different algorithms, e.g., ‘stumps’ in AdaBoost, leaf-wise growing in Light GBM, symmetric ones in CatBoost. In particular, gradient boosting uses gradient descent for adjusting the weights and XGBoost provides an improved version which runs much faster, handles missing values on its own and defies the risk of overfitting to a large extent.

Finally, once we have trained a number of ensemble (or other) models, and ranked their performance according to a proper evaluation metric, we can obtain our predictions based on all of them, through a process named ‘stacking’. That is yet another weighted sum, this time on a higher level of hierarchy, which AutoGluon does automatically in `WightedEnsemble_L2`. The idea is again that different approaches of decision making balance each other and thus a lot of their imperfections are being cancelled out. Our experiments confirm this expectation as the ensemble method wins in all scenarios, except for one particular case of very biased data which we might discard from the very beginning.

2.3. Optimization and Evaluation Metrics

Complex algorithms like Random Forest and XGBoost that we use here, have additional intrinsic parameters (hyperparameters), such as the learning rate or the maximal depth of trees, whose optimal values differ according to the specific context. To make it even worse, some of these parameters are categorical, e.g., the type of activation function in classification problems. Hence, hyperparameter tuning is a highly non-trivial problem in of itself and requires lots of computational power to determine how the default values need to be altered in each particular case in order to benefit the overall performance of the model, so instead of brute-forcing with GridSearch, we use RandomizedSearch which runs much faster at the expense of allowing some uncertainty. The evaluation is being done via k-fold cross-validation, which means the training-test data split is performed k times and the corresponding metric values are then being averaged for the overall outcome. This way we avoid false evaluation due to biased data in a specific sample. As for the evaluation metrics, for regression problems it is customary to use the so-called coefficient

of determination r^2 , equal to the square of the Pearson correlation coefficient. Another commonly used predictor of accuracy is the mean absolute percentage error MAPE, i.e., the average of the absolute values of all relative errors in the predictions. What we refer to as ‘accuracy’ here is simply $1 - \text{MAPE} [\%]$, evaluated using cross-validation (just like r^2).

3. Materials and Methods

We tackle the problem with the overwhelming amount of missing traffic data (in some cases more than 99.5%) by using a wide variety of features, grouped in several categories:

- infrastructure including the street class, capacity, pavement, number of directions etc.
- space syntax: integration and choice parameters with averaged distance
- functional analysis: points of interest and cadastral built-up area data modeled as concentration of activities and motorized users
- demographics: density of motorized inhabitants
- measured data: several traffic counts on primary and secondary streets and national roads, NO₂ levels.

Some of the features we use are already a result of modeling, e.g., the Open Transport Map (OTM) traffic model, as well as interpolations of different results from previous studies. Table 1 describes briefly the different features we use as predictors in our model (cf [18,19]).

Table 1. Features used for training the ML model.

Column Name	Data Content	Values
baseTYPE	street type classification	40,736
IMMIS_RT	traffic situation typology	7637
C_KAPAZ	estimated road capacity	7637
StrDMNDRCT	number of directions	40,736
EMIT_SPEED	speed limit	7637
EMIT_GRDNT	street slope (gradient)	7637
SSINTr_In	space syntax (integration)	40,736
SSCHr_In	space syntax (choice)	40,736
X0, Y0	coordinates of the centroids	40,736
OTMsurface	Open Transport Map (OTM) street surface type	5710
OTMtraffic	OTM traffic model	7637
TT200mHMc	TomTom traffic count data heat map r = 200 m	7637
ACTUSEmean	heatmap r = 200 m estimated motorized users POI and cadastral data based	40,651
ACTLIVmean	heatmap r = 200 m estimated motorized inhabitants census based	40,497
exIDWmean	IDW-interpolated point-based RF traffic model ¹	7637

¹ We used clusterization into segments according to traffic and consecutive data imputation with RF-regression.

These are quite different types of data, as is their role in our model. The first few groups are categorical and we encode them as dummy variables. Together with the space syntax and demographic features they can be used for clustering of the data set which may improve the prediction accuracy significantly as we have seen in our previous studies. However, some of the clusters in this particular case end up with too few values to train and test the model, thus we only use the street type and distinguish between primary (type A) and secondary (type B) urban street network. Type A (7637 rows of data) consists of major traffic arteries, including city highways and boulevards, while type B (33,099 rows) includes secondary streets with much smaller traffic capacity. The geographic coordinates, which also play important role in our model, correspond to the centroids of the street segments described by each row in the data set. Naturally, the measurements of population density, activities, pollution etc. have not been conducted exactly in those points, so we use heat map and inverse distance weighted (IDW) interpolation or nearest joins to obtain a good estimate of those values in the corresponding row. As for the traffic count, as it is usually conducted at crossroads, we also need such techniques to evaluate it for street segments and hence to their centroids, by averaging. We rely on different sources for the different years—some provide an estimate for the daily or yearly average, others—time series for specific days of the year. We also conducted our own counts in certain points of the urban network, and estimated the corresponding daily average using the temporal dynamics

for the ‘nearest’ (with respect to the distance matrix) location, for which it is available. As might be expected, the model for the secondary streets feeds on much less features as less measurements are available for this neglected part of the urban infrastructure. However, it relies on a much larger number of training rows. The correlations are illustrated with heat maps in Figure 1. We also note that the problem of multicollinearity does not bother us since we use ensemble learning algorithms that are unaffected by it. For the same reason no data normalization has been performed—the only preprocessing we did was cleaning.

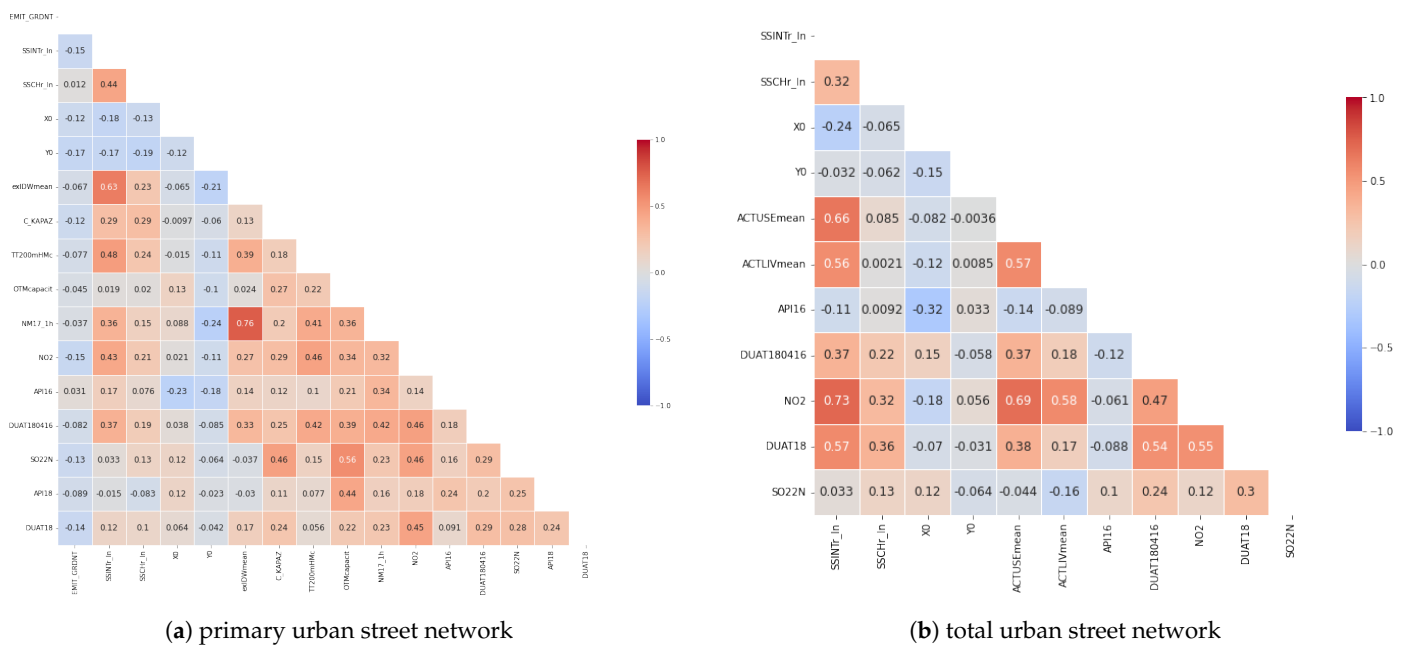


Figure 1. Correlation heat map for the features and target variables used in the model.

4. Results

Next, we present the experimental results of our manipulation with the data and their interpretations, starting with the technical parameters of the machine learning algorithms.

4.1. The Primary Street Network

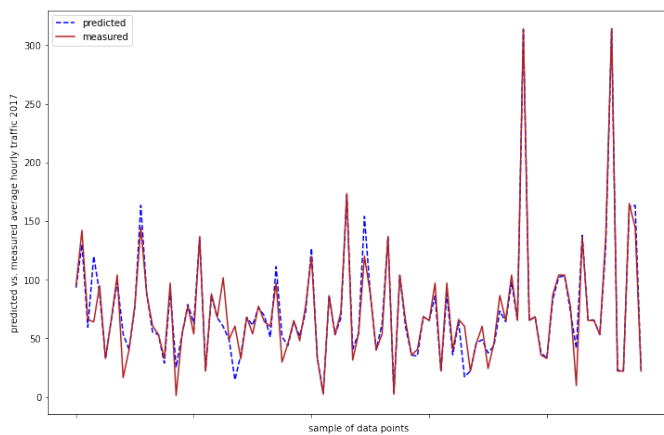
Our strategy is to build up to the traffic model gradually, starting with data imputation for the features that are useful in the training process and cannot be reliably interpolated. Note that there are two different data rows for the average daily traffic in 2018 (the third one is a mix) as there are two sources of measurements. ‘API18’ shows clear signs of overfitting—the reason being not the selected algorithm but the data bias: all measurements have been done in strongly correlated segments from the major traffic arteries, near the city rings. Hence, we dismiss this prediction, as well as the one for the merged columns, considering it to be infested by this bias (‘API16’ has a similar problem). Other measurements cover only small time intervals and the corresponding daily averages have been evaluated using the temporal dynamics for the nearest (with respect to the origin-destination matrix) network vertex for which such data is available. All this naturally amounts to an error that is difficult to estimate. So, we have one type of data that is time-averaged but lacks sufficient spatial coverage (‘API’) and another one, much better put coordinate-wise while based on time-limited measurements (‘DUAT’). The situation with nitrogen dioxide levels resembles the latter while ‘SO22N’ mixes different sources and is a compromise between the two. As our study is focused on spatial distribution, its natural target variables are ‘DUAT18’, ‘SO22N’ and ‘NO₂’, while we treat other features merely as an aiding tool for the models.

Table 2 above shows the accuracy of ensemble learning regression algorithms used for data imputation, while Figure 2 illustrates how our models perform on the primary street network when sufficient amount of training data is available and when that is not the case.

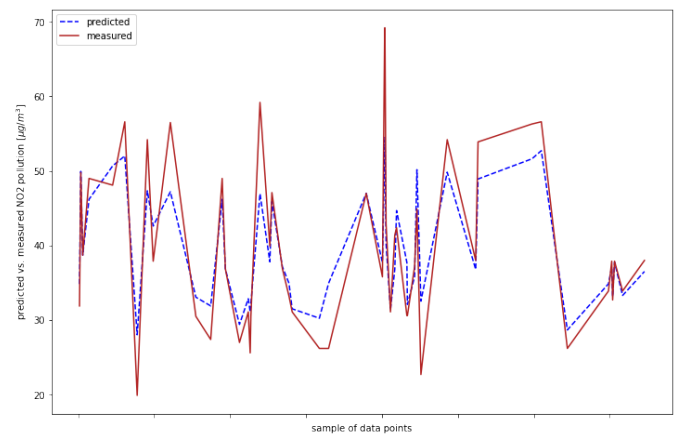
Table 2. Features with missing values predicted by the corresponding model ¹.

Feature	Content	% Missing	Model	Accuracy	r ²
OTMcapacit	capacity	25.23%	XGB/RSO	82.62%	0.83
NM17_1h	hourly 2017 traffic count	93.70%	WE_L2	64.29%	0.94
NO ₂	NO ₂ levels	99.21%	WE_L2	90.04%	0.74
API16	road traffic	92.61%	XGB/RSO	92.60%	0.91
DUAT180416	street traffic	60.05%	WE_L2	80.02%	0.43
SO22N	street traffic	99.36%	WE_L2	78.84%	0.33
API18	bidirectional road traffic	99.63%	WE_L2	81.33%	0.97
MIX18	mixed traffic	97.81%	WE_L2	71.68%	0.45
DUAT18	street traffic	98.18%	WE_L2	67.95%	0.25

¹ Ranked by accuracy based on MAPE (mean absolute percentage error) estimates. We also use the abbreviations: RSO (Random Search Optimization), WE_L2 (Weighted Ensemble_L2), XGB (Extreme Gradient Boosting).



(a) hourly traffic count in 2017



(b) average NO₂ levels in 2021

Figure 2. Predictions and test data for different features.

It is worth mentioning also the feature importance of different parameters. We show two examples in Figure 3. This marker shows how each feature affects the performance and accessing it allows us to understand the models better and improve them dynamically.

Table 3 below demonstrates how AutoGluon ranks the best performing algorithms with respect to a chosen evaluation metric, in this case the coefficient of determination is r². The values are obtained via cross validation and the ranking using MAE or MAPE is similar. The tables for all modeled features look quite alike. In some cases instead of AutoGluon we use the standard RF or XGB regressors and possibly fine parameter tuning with random search optimization, as shown in Table 2 as well as Table 4 in the next section.

Table 3. Best performing models for the NM17_1h data imputation task according to AutoGluon ¹.

Model	Score_Val	Fit_Time	Fit_Order
WeightedEnsemble_L2	0.894	23.48	9
RandomForestMSE	0.881	0.715	3
XGBoost	0.881	1.639	7
NeuralNetTorch	0.863	20.33	8
ExtraTreesMSE	0.848	0.637	5
CatBoost	0.763	3.560	4
NeuralNetFastAI	0.576	1.354	6

¹ The coefficient of determination r² calculated via cross-validation is used as evaluation metric (score_val).

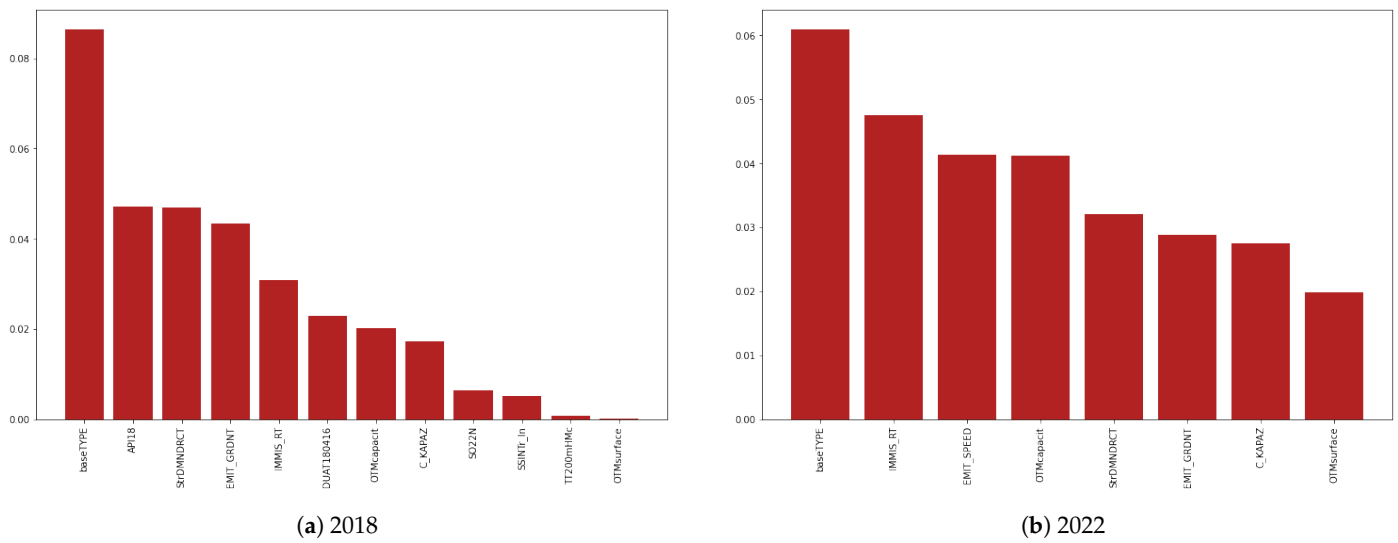


Figure 3. Feature importance in experimental daily street traffic models for two different years.

4.2. The Entire Street Network

Unfortunately, methodologically sound gathering of traffic data for cities like Sofia is still a luxury at this point. We have to settle with the latter case and this is the reason behind our choice to use as input data an array of categorizations, direct entry or nearest joins of attribute records from the strategic noise map as well as NO₂ levels (which were later processed by our ML model), IDW interpolation of point-based RF traffic model data for the junctions of the primary street and road network and heat map techniques with radius of 200 m providing continuity and granularity of uneven data, e.g., from sample based traffic counts of TomTom and functional analysis of points of interest or address points with assumed motorization rates of users or inhabitants. As for the secondary streets we have too few measurements, we use our own randomly sampled short-term observations from 2021 at the city scale and for 2022 in the extents of official low emission zones (cf [20]).

Figure 4 illustrates two features that become more important from the model as we go farther from the main streets, and have to do more with demographics and urban dynamics, while on Figure 5 we also show the predictions of our traffic ML models for 2018 and 2022.

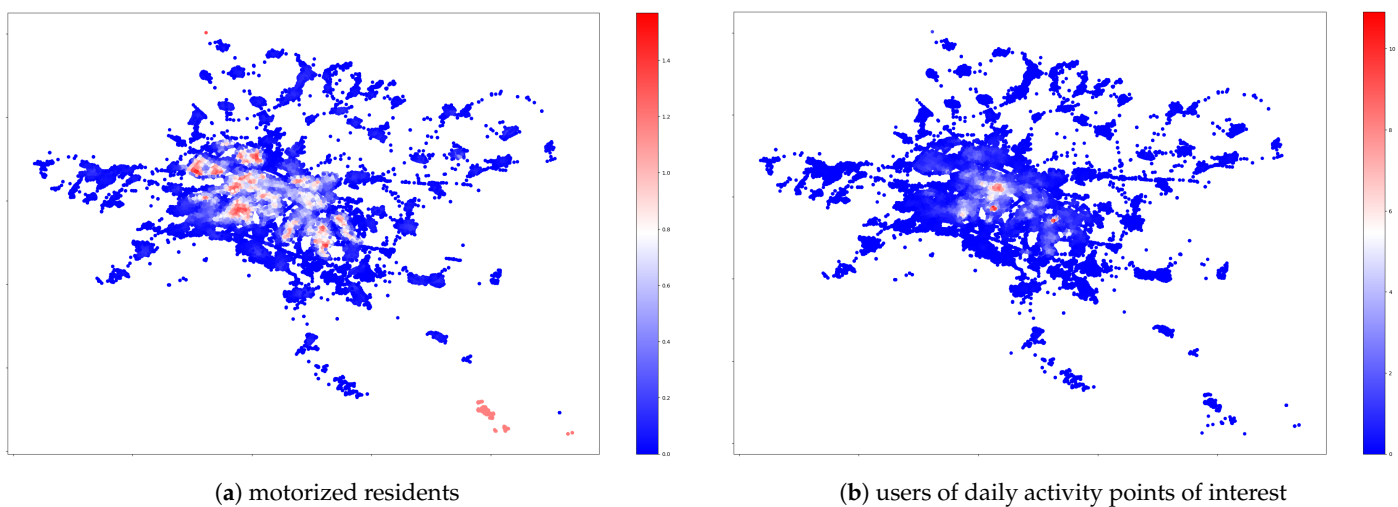


Figure 4. Distribution of inhabitants (measured in thousands) used in urban functional analysis.

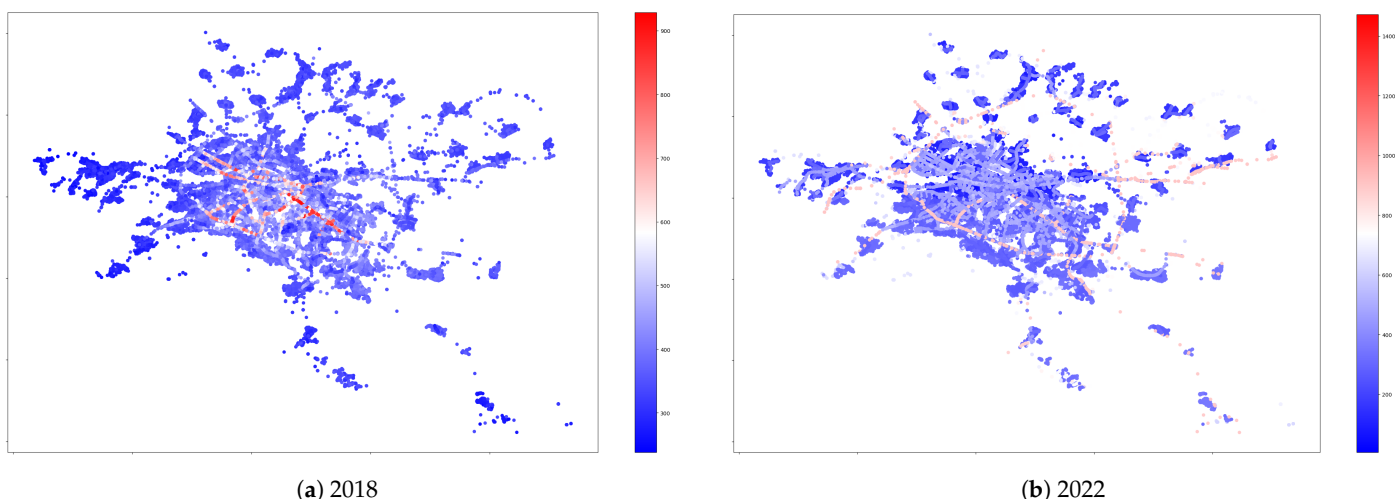


Figure 5. Scatter plot of the average hourly traffic for different years predicted with AutoGluon.

Let us point out that the role of different hyperparameters remains somewhat obscure in the weighted ensemble as various different models take place in the stacking process. For this reason we do not pay special attention to hyperparameter tuning (it is a standard part of the AutoML embedded optimization). Note that we use single model predictions (either RF or XGB) only to impute missing data for some of the intermediate features, and not the actual target variables. Even if we use fine tuning (RSO), the predictive power of the features, e.g., the road traffic in the 'API' and 'NM17' categories, is relatively low and they only 'balance' the model. Despite the reasonably high accuracy for these predictions, we do not consider them to be a reliable end product of our study, the reason being that the data samples are far from representative for the urban traffic when restricted only to the intercity roads or separate hours of counting in different days. 'DUAT', 'SO' and 'NO₂' measurements on the other hand have this property, being more comprehensive, so the only results of scientific and practical value we claim here are for these target variables.

Table 4. Features with missing values predicted for the whole city network (see [19–24] for raw data).

Column ¹	Content	Missing Data	Best Model	Accuracy	r ²
ACTUSEmean	active users of functions	0.21%	RF	97.36%	0.97
ACTLIVmean	density of inhabitants	0.59%	RF	97.88%	0.98
API16	road traffic	98.62%	RF	93.57%	0.90
DUAT180416	street traffic	92.51%	WE_L2	88.83%	0.69
NM17_1h	hourly noise map traffic	99.82%	WE_L2	84.10%	0.91
NO ₂	NO ₂ levels	99.85%	WE_L2	91.90%	0.80
API18	road traffic	99.93%	XGB/RSO	90.92%	0.61
DUAT18	street traffic	81.25%	WE_L2	77.78%	0.24
SO22CC	street traffic	99.79%	WE_L2	69.12%	0.67
DUAT18/WE_L2	modeled	81.25%	WE_L2	95.06%	0.65
SO22CC/WE_L2	modeled	81.16%	WE_L2	94.22%	0.92

¹ The last two rows refer to already modeled data (Table 2), while 'SO22CC' includes also some small streets.

5. Discussion

The present piece of research is a part of a broader research project focused on urban development scenarios, mobility, air quality and health, particularly on the path to more integrated modeling methodology development and in the context of digital transition priority themes in the European Union, Bulgaria and Sofia municipality [25–27]. Our effort was preceded by other modeling experiments [20,28], the latter being used for air pollution dispersion modeling and simulations [29]. They illustrate the valuable aid of various machine learning methods in revealing traffic patterns based on quantitative and qualitative

data. The main question we pose is “What are the optimal ML algorithms for urban and peri-urban traffic modeling using various indirect features, some including incomplete and/or low resolution data?”. The obvious application is a commonly encountered setting, in which data gathering or access practices don’t work smoothly enough to allow for good governance, planning and management of transportation, optimal development and healthy urban environment, leading to health issues and seriously reduced quality of life for the urban population of numerous cities. Therefore, at the EU level there are several legislative and political frameworks that support the implementation of new approaches, methods and techniques, which can benefit from the ML advantages and ML techniques can be performed in more consistent way, namely the intelligent transport systems [30], the sustainable and smart mobility [31], the current and future air quality modeling requirements [32], the access to public information [33] and the infrastructure for spatial information [34]. Most of these policies are either transposed in the legislation or officially supported at the national level of EU member states. Various national or regional agencies and local authorities adhere to the practices of decentralized management and accessibility of data. In many instances, however, expert and institutional environments are not built on mutual trust but on opportunistic distribution of influence leading to separate campaigns of provision of public data, especially in the domain of traffic, while there are cities and whole regions with severe environmental burden caused by air pollution, plus other transport related problems, such as noise, whose abatement seems really slow [35].

6. Conclusions

The present study focuses on the problem of missing traffic data values for cities with underdeveloped informational infrastructure, choosing Sofia as a good example. Such cities are often neglected by researchers as it is very hard for them to find reliable data in order to feed their models. However, the problems with traffic, pollution and health issues are usually worse namely in those cities, so it is worth putting some extra effort. Sofia is by no means too much lagging behind capital city, but falls behind in terms of data gathering and sharing, 5G and IoT infrastructure, data warehouses, etc., like many other cities in Eastern Europe and globally with similar problems. Here we focus mainly on the overall traffic estimates, using pollution only as a predictor, but this study is just a part of a broader research project aiming to tackle environmental and health issues in Sofia, for which urban traffic seems to play a crucial role in the air pollution emissions generation counterpart. Based on the results presented in this paper, it seems that our faith in ensemble learning algorithms has been justified. They outperform other standard machine learning and statistical tools in most urban traffic modeling settings—quite a ‘muddy’ area of research by accuracy standards. Nevertheless, even with a relatively small sample of reliable data, covering a diverse enough set of features sufficiently correlated with the target variable, it is still possible to obtain reasonable prediction, as we demonstrate here for the city of Sofia. While still waiting for better data management and sharing environment that could feed our future models more abundantly, we continue to dedicate efforts to optimise and test the our current findings. We dig deeper in the available data for the reanalysis of air pollution measurements, in recent years increasingly linked to transportation activities, but also in relation to noise and access to green areas. We consider the present study a step towards building plausible scenarios for the near future urban development, environment, mobility and health related policies based on modeling the mitigation for the negative effects of traffic and other air pollution sources thanks to the combination of wise transport and green infrastructure development.

Author Contributions: Conceptualization, A.B. and D.B.; methodology, D.B. and A.B.; software, D.B. and A.B.; validation, D.B. and A.B.; formal analysis, D.B.; investigation, D.B. and A.B.; resources, A.B.; data curation, A.B.; writing original draft preparation, D.B.; writing review and editing, D.B. and A.B.; visualization, D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Bulgarian National Science Found (BNSF), under project “Development of a methodology for assessing air quality and its impact on human health in an urban environment”, grant number: KP-06-H54/2, 15 November 2021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We provide the raw data along with the results of our models, respectively at https://docs.google.com/spreadsheets/d/17CtZoDBpatKS5OkJ3PxNj7cRy_1q9Hi3s3BiTNWUz8/edit?usp=sharing (accessed on 27 February 2023) and https://docs.google.com/spreadsheets/d/1Q_fSBf1OWtW5b6yU3e8Oj9fw9J_qj-TN-GtoW-cKAA4/edit?usp=sharing (accessed on 27 February 2023) for the primary streets, and https://docs.google.com/spreadsheets/d/1JskUDwFSPeAIR_JYjvtgtoVBzKwdhZ-0cU_rtNins-A/edit?usp=sharing (accessed on 27 February 2023) for the entire urban network.

Acknowledgments: We are grateful to the environmental NGO ‘ZaZemiata’ for initial steps of support around early testing of methods and modeling which then evolved thanks to the funding from the BNSF and also for providing us and the general public with data about the NO₂ pollution in Sofia which seems to be of great value thanks to the undertaken citizens science effort.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MAPE	Mean average percentage error
IDW	Inverse Distance Weighted
RF	Random Forest
RSO	Random Search Optimization
XGB	Extreme Gradient Boosting
WE_L2	WeightedEnsemble_L2

References

1. Lohrasbinasab, I.; Shahraki, A.; Taherkordi, A.; Delia Jurcut, A. From statistical- to machine learning-based network traffic prediction. *Trans. Emerg. Tel. Tech.* **2020**, *33*, e4394. [CrossRef]
2. Alqudah, N.; Yaseen, Q. Machine Learning for Traffic Analysis: A Review. *Procedia Comput. Sci.* **2020**, *170*, 911–916. [CrossRef]
3. Zhang, Y.; Liu, J.; Shen, W. A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. *Appl. Sci.* **2022**, *12*, 8654. [CrossRef]
4. Mahdavian, A.; Shojaei, A.; Salem, M.; Laman, H.; Yuan, J.-S.; Oloufa, A. Automated Machine Learning Pipeline for Traffic Count Prediction. *Modeling* **2021**, *2*, 482–513. [CrossRef]
5. Erickson, N.; Shi, X.; Sharpnack, J.; Smola, A. Multimodal AutoML for Image, Text and Tabular Data. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), Washington, DC, USA, 14–18 August 2022; pp. 4786–4787. [CrossRef]
6. Du, S.; Li, T.; Gong, X.; Horng, S. A Hybrid Method for Traffic Flow Forecasting Using Multimodal Deep Learning. *Int. J. Comput. Intell. Syst.* **2020**, *13*, 85–97. [CrossRef]
7. Sfyridis, A.; Agnolucci, P. Annual average daily traffic estimation in England and Wales: An application of clustering and regression modeling. *J. Transp. Geogr.* **2020**, *83*, 102658. [CrossRef]
8. Pun, L.; Zhao, P.; Liu, X. A Multiple Regression Approach for Traffic Flow Estimation. *IEEE Access* **2019**, *7*, 35998–36009. [CrossRef]
9. Fathurrahman, M.F.; Sutarto, H.Y.; Semajnski, I. Urban Network Traffic Analysis, Data Imputation, and Flow Prediction based on Probabilistic PCA Model of Traffic Volume Data. In Proceedings of the 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Bandung, Indonesia, 29–30 September 2021; pp. 1–6. [CrossRef]
10. Joelianto, E.; Fathurrahman, M.F.; Sutarto, H.Y.; Semajnski, I.; Putri, A.; Gautama, S. Analysis of Spatiotemporal Data Imputation Methods for Traffic Flow Data in Urban Networks. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 310. [CrossRef]
11. Jayasinghe, A.; Sano, K.; Abenayake, C.C.; Mahanama, P.K.S. A novel approach to model traffic on road segments of large-scale urban road networks. *MethodsX* **2019**, *6*, 1147–1163. [CrossRef] [PubMed]

12. Zhao, S.X.; Wu, H.W.; Liu, C.R. Traffic flow prediction based on optimized hidden Markov model. *J. Phys. Conf. Ser.* **2019**, *1168*. [CrossRef]
13. Feng, B.; Xu, J.; Zhang, Y.; Lin, Y. Multi-Step Traffic Speed Prediction Based on Ensemble Learning on an Urban Road Network. *Appl. Sci.* **2021**, *11*, 4423. [CrossRef]
14. Bokaba, T.; Doorsamy, W.; Paul, B.S. A Comparative Study of Ensemble Models for Predicting Road Traffic Congestion. *Appl. Sci.* **2022**, *12*, 1337. [CrossRef]
15. Khan, N.U.; Shah, M.A.; Maple, C.; Ahmed, E.; Asghar, N. Traffic Flow Prediction: An Intelligent Scheme for Forecasting Traffic Flow Using Air Pollution Data in Smart Cities with Bagging Ensemble. *Sustainability* **2022**, *14*, 4164. [CrossRef]
16. Siemens. European Green City Index. 2009. Available online: <https://assets.new.siemens.com/siemens/assets/api/uuid:fdde99e7-5907-49aa-92c4-610c0801659e/european-green-city-index.pdf> (accessed on 27 February 2023).
17. Phrenos. Expert Panel Technical Assessment Synopsis Report European Green Capital Award 2023. Available online: https://ec.europa.eu/environment/europeangreencapital/wp-content/uploads/2021/07/EGCA_2023_Technical_Assessment_Synopsis_Report.pdf (accessed on 27 February 2023).
18. Jedlička, K.; Ježek, J.; Kolovský, F.; Kozhukh, D.; Martolos, J.; Šťastný, J.; Charvát, K.; Hájek, P.; Beran, D. Open Transport Map. 2015. Available online: <https://opentransportmap.info/> (accessed on 27 February 2023).
19. Sofiaplan Open Data (in Bulgarian). 2022. Available online: <https://api.sofiaplan.bg> (accessed on 27 February 2023).
20. Za Zemiata. Spatially Based Scenarios for Introduction of Low Emission Zones in Stolichna Municipality (in Bulgarian). 2023. Available online: <https://www.zazemiata.org/resources/report-transport-lez-sofia/> (accessed on 27 February 2023).
21. Spektri EOOD; GIS Sofia. Development of an updated Strategic Environmental Noise Map of the Sofia agglomeration (in Bulgarian). 2017. Available online: https://www.sofia.bg/documents/20182/3044533/2018-05-14-Sofia_ShKarta2017_ObedineniDoc.pdf/915739c6-3876-439f-a871-1421736efd2d (accessed on 27 February 2023).
22. Za Zemyata. Is There Air Pollution in Sofia with Nitrogen Dioxide? 2021 Za Zemyata Measurement Results (in Bulgarian). 2022. Available online: <https://www.zazemiata.org/wp-content/uploads/2022/05/Za-Zemyata-Doklad-NO2-Online.pdf> (accessed on 27 February 2023).
23. Burov, A. Counting Traffic on Small Streets in the City of Sofia 07.2021 (in Bulgarian). 2022. Available online: <https://bpos.bg/publication/33973> (accessed on 27 February 2023).
24. INNOAIR. White Paper on the Introduction and Effective Operation of Low-emission Zones for Motor Vehicles on the Territory of the Metropolitan Municipality (in Bulgarian). 2022. Available online: https://www.innoair-sofia.eu/images/documents/documents-bg/04_2_White_Book_V4_m.pdf (accessed on 27 February 2023).
25. European Commission. Urban Agenda for the EU. 2023. Available online: https://commission.europa.eu/eu-regional-and-urban-development/topics/cities-and-urban-development/urban-agenda-eu_en (accessed on 27 February 2023).
26. National Program for Atmospheric Air Quality Improvement (in Bulgarian). 2019. Available online: <https://www.mtc.government.bg/bg/category/42/integrirana-transportna-strategiya-v-perioda-do-2030-g> (accessed on 27 February 2023).
27. Stolichna Municipality. Strategy for Digital Transformation of Sofia. 2020. Available online: <https://www.sofia.bg/w/strategia-za-digitalna-transformacia-na-sof-1> (accessed on 27 February 2023).
28. Burov, A.; Brezov, D. Transport Emissions from Sofia's Streets—Inventory, Scenarios, and Exposure Setting. In *Environmental Protection and Disaster Risks*; Dobrinkova, N., Nikolov, O., Eds.; EnviroRISKs 2022: Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, 2023; Volume 638.
29. Velizarova, M.; Dimitrova, R. Impact of Regulatory Measures on Pollutants Concentration in Urban Street Canyon—A Pilot Study. In *Environmental Protection and Disaster Risks*; Dobrinkova, N., Nikolov, O., Eds.; EnviroRISKs 2022: Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, 2023; Volume 638.
30. Directive 2010/40/EU of the European Parliament and of the Council of 7 July 2010 on the Framework for the Deployment of Intelligent Transport Systems in the Field of Road Transport and for Interfaces with Other Modes of Transport. Available online: <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32010L0040> (accessed on 27 February 2023).
31. European Commission. Sustainable and Smart Mobility Strategy—Putting European Transport on Track for the Future. 2020. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0789> (accessed on 27 February 2023).
32. Proposal for a Directive of the European Parliament and of the Council on Ambient Air Quality and Cleaner Air for Europe (Recast) COM/2022/542 Final. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A542%3AFIN> (accessed on 27 February 2023).
33. Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on Public Access to Environmental Information and Repealing Council Directive 90/313/EEC. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32003L0004> (accessed on 27 February 2023).

34. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Available online: <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32007L0002> (accessed on 27 February 2023).
35. EEA. The European Environment—State and Outlook 2020. Knowledge for Transition to a Sustainable Europe. 2019. Available online: <https://www.eea.europa.eu/soer/publications/soer-2020> (accessed on 27 February 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.