sustainability

MDPI

Article

Managing Traffic Data through Clustering and Radial Basis Functions

Heber Hernández [1], Elisabete Alberdi [2,*], Heriberto Pérez-Acebo [3], Irantzu Álvarez [4], María José García [4], Isabel Eguia [2] and Kevin Fernández [2]

1 Nube Minera, La Serena 1700000, Chile; heber@nubeminera.cl
2 Department of Applied Mathematics, University of the Basque Country UPV/EHU, 48013 Bilbao, Spain; isabel.eguia@ehu.eus (I.E.); kevin_fer_martinez@hotmail.com (K.F.)
3 Mechanical Engineering Department, University of the Basque Country UPV/EHU, 48013 Bilbao, Spain; heriberto.perez@ehu.eus
4 Department of Graphical Expression and Engineering Projects, University of the Basque Country UPV/EHU, 48013 Bilbao, Spain; irantzu.alvarez@ehu.eus (I.Á.); mariajose.garcialopez@ehu.eus (M.J.G.)
* Correspondence: elisabete.alberdi@ehu.eus; Tel.: +34-946-017-790

Abstract: Due to the importance of road transport an adequate identification of the various road network levels is necessary for an efficient and sustainable management of the road infrastructure. Additionally, traffic values are key data for any pavement management system. In this work traffic volume data of 2019 in the Basque Autonomous Community (Spain) were analyzed and modeled. Having a multidimensional sample, the average annual daily traffic (AADT) was considered as the main variable of interest, which is used in many areas of the road network management. First, an exploratory analysis was performed, from which descriptive statistical information was obtained continuing with the clustering by various variables in order to standardize its behavior by translation. In a second stage, the variable of interest was estimated in the entire road network of the studied country using linear-based radial basis functions (RBFs). The estimated model was compared with the sample statistically, evaluating the estimation using cross-validation and highest-traffic sectors are defined. From the analysis, it was observed that the clustering analysis is useful for identifying the real importance of each road segment, as a function of the real traffic volume and not based on other criteria. It was also observed that interpolation methods based on linear-type radial basis functions (RBF) can be used as a preliminary method to estimate the AADT.

Keywords: average annual daily traffic; sustainable management; spatial analysis; RBFs; clustering; road network level

Citation: Hernández, H.; Alberdi, E.; Pérez-Acebo, H.; Álvarez, I.; Garcia, M.J.; Eguia, I.; Fernández, K. Managing Traffic Data through Clustering and Radial Basis Functions. *Sustainability* **2021**, *13*, 2846. https://doi.org/10.3390/su13052846

Academic Editor: Armando Cartenì

Received: 19 October 2020
Accepted: 25 February 2021
Published: 5 March 2021

1. Introduction

Since the second half of the 20th century, road transport has become the most important transport mode in Europe, with a quote over 75% of the transport modal distribution, for both passengers and freight. This trend is almost identically repeated in all the countries of the European Union [1]. Nevertheless, this road traffic volume is not equally distributed between road network levels, and, generally, the most important level is used by the majority of the road traffic, approximately 50% of road traffic volumes, once again for both passenger and freight traffic, although this top road level comprises a small portion of the total length of the road network. This trend can be observed in any developed country [2–4].

The necessity of dividing the road network of a state, region, or province in various levels is subjected to the functions that the road network have to fulfill. These functions are mobility and accessibility. Mobility is the property that evaluates the number and quality of the trips, taking into account the traffic volume and the speed or travel times. A quick, comfortable and safe circulation is aimed to provide. On the other hand, the accessibility is the property that measures the easiness to access any part of the territory, allowing to reach

any inhabited location. These functions are competing objectives. Whereas in a road that prioritize mobility impediments to the traffic flow should be minimized, roads providing access to adjacent land uses have more frequent access points [2,3]. The compromise between mobility and accessibility influences the operation and the safety of the roads; hence, the network or any road should be planned carefully, according to their function or functions. Each type of road has a vital role for developing a well-operating and efficient network with the aim of facilitating higher-speeds in long-distance travels and lower speeds in short-distance trips. In general terms, roads can be classified in arterials, collectors, and local roads, with a variable proportion of the functions (mobility and access); see Figure 1. Arterials represent the highest level of mobility and are focused on providing high-speed and uninterrupted flow. Long-distance travels are frequent in arterials. Collectors have a blended objective of maintaining mobility and access, facilitating the movement between local roads and arterials. Finally, local roads or streets provide direct access to residences, businesses, and other land uses. On local roads, speed can be minimized to provide more frequent accesses to adjacent lands [3].

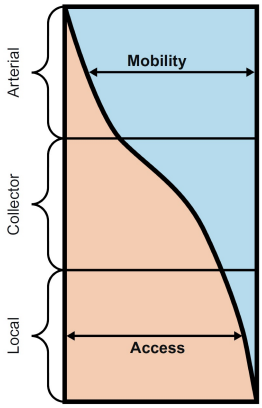

Figure 1. Relationship of functionally classified road network levels according to the fulfillment of the functions of mobility and accessibility.

Moreover, apart from the functional classification of roads, more inputs are needed in any transport planning. A transport planning, at any level (city, regional, or statewide level) includes a large transport network, various stakeholder groups, a wide range of land uses and their influence, and an elaborated and accurate planning process [5]. Usually, four steps are present in a transport planning: trip generation, trip distribution, mode choice, and trip assignment. The final input of this process is a travel demand forecasting model. With regard to the first step, trip generation, it implies the collection and analysis of data about transport, such as traffic volumes, population, housing, type and rate of employment, and vehicle ownership, for simulating existing travel at present and future travel pattern [2,5]. For measuring traffic volumes in highways, which was identified as the main transport mode, traffic counts are employed. They can be temporary or permanent. The permanent count stations are necessary to know the variability in the traffic volumes, which is observed between seasons, months of the year, days of the week, hours of the day, and even during short-term periods below an hour [6]. The most common output from a count station is the average annual daily traffic (AADT), which can be calculated dividing by 365 the total number of vehicles passing a given point during a year. AADT data are generally published by highway authorities in an annual report, indicating the data for each of the studied roadways in the network under their management [2,5].

The average annual daily traffic is an essential parameter in many aspects of the highway engineering, apart from comparing the importance of the roads in a network. If hourly volumes are calculated, these data are used for features of the geometric design of roads. As expected, traffic volumes have impact on road safety [7–12]. Additionally, AADT is also introduced in pavement deterioration models as a key factor, as it represents

the quantity of loads that pavements must withstand [13]. For parameters, like pavement roughness or pavement condition, expressed by means of indices, such as the International Roughness Index (IRI), the Present Serviceability Index (PSI), or the Pavement Condition Index (PCI), the AADT or the cumulated volumes during a period are employed. Hence, the ADDT of each year must be known [14–20]. However, for other characteristics, such as the skid resistance of pavements, the AADT is the key element in the modeling because the available friction varies according to the traffic volume of each year [21–25]. Furthermore, the AADT has been also correlated with other fields, such as the air-pollution and noise [26–28] and bridge deterioration [29].

Nonetheless, traffic counts cannot be placed on every road section of all the roads due to various reasons [30] and various approaches have been proposed in the literature to deal with this problem of lack of data. For example, Sfyridis and Agnolucci [31] proposed a methodology to predict AADT by means of clustering and regression modeling with variables about roadway, socioeconomic, and land use characteristics. Similarly, Wu and Xu [32] presented a multiple regression model with logarithmic transmission for AADT predictions for minor roads at intersections after comparing various possible models. Ma et al. [33] developed a copula-based model for forecasting AADT, which can describe spatial dependency and is robust to outliers. Artificial Neural Networks and Deep Learning have also been employed to analyze the accuracy of various neural network-based models for estimating AADT and hourly traffic volumes [34,35]. Finally, Chang and Cheon [36] underlined the strong relationship between AADT and vehicle GPS data, which can be used for unmeasured road segments.

In this work, a study of the AADT is performed using clustering and spatial methods, and data from the count stations located in the Basque Autonomous Community (BAC) in Spain as a case study. The aim is to perform a precise analysis of the traffic volume of the roads in the BAC and, subsequently, to be able to create a map of the BAC depending on the traffic-volume using the radial basis function interpolation. This analysis can be used in posterior studies as a preliminary step to plan and build new roads in areas in which there is no information, or even to adjust an adequate maintenance method to the existing infrastructures with the purpose of ensuring the sustainability of the roads and different issues related to them.

The article is organized as follows: in Section 2, first, the studied area and the data used for the analysis are presented, and second, the theoretical framework on which the calculations are based is exposed; in Section 3, the exploratory data analysis is done; in Section 4, main results obtained both in the clustering analysis and in the estimation model and its validation are presented. Finally, in Section 5, conclusions derived from the study are reported.

2. Methodology

2.1. Studied Area and Data

The Basque Autonomous Community is located in the northern part of Spain. It is composed by three provinces (Biscay, Gipuzkoa, and Álava), and it covers 7230 km^2. It has 2.178 million inhabitants, of which 40% are concentrated in the Bilbao Metropolitan Area [37], which is also the area with most industry and most traffic. Each of the provinces have the ownership and competences over all the highways in its territory, including the main freeways and highways that communicate with other regions of Spain, except from the municipal streets [19,25]. The road agencies of each of the Regional Governments manages its own road networks and publishes annually the traffic data. In this research, those corresponding to 2019 were employed [38–40].

Measurement data used herein are: the Annual Average Daily Traffic (AADT), percentage of heavy traffic, number of lanes in each direction, speed limit in the segment, name of the road, network level which the road belongs to, and the type of count station.

With regard to the functional classifications of roads, a very similar organization is used in the three provinces; see Figure 2 and Table 1. Roads are divided in the preferential

interest network (red network), the basic network (orange), the provincial network (green), and the local network (yellow), as presented in Table 1. Additionally, there is the complementary network (blue) in Biscay, between the basic and the provincial one, which appears only in the metropolitan area of Bilbao; and a neighborhood network in Álava, which is the least important network (grey).

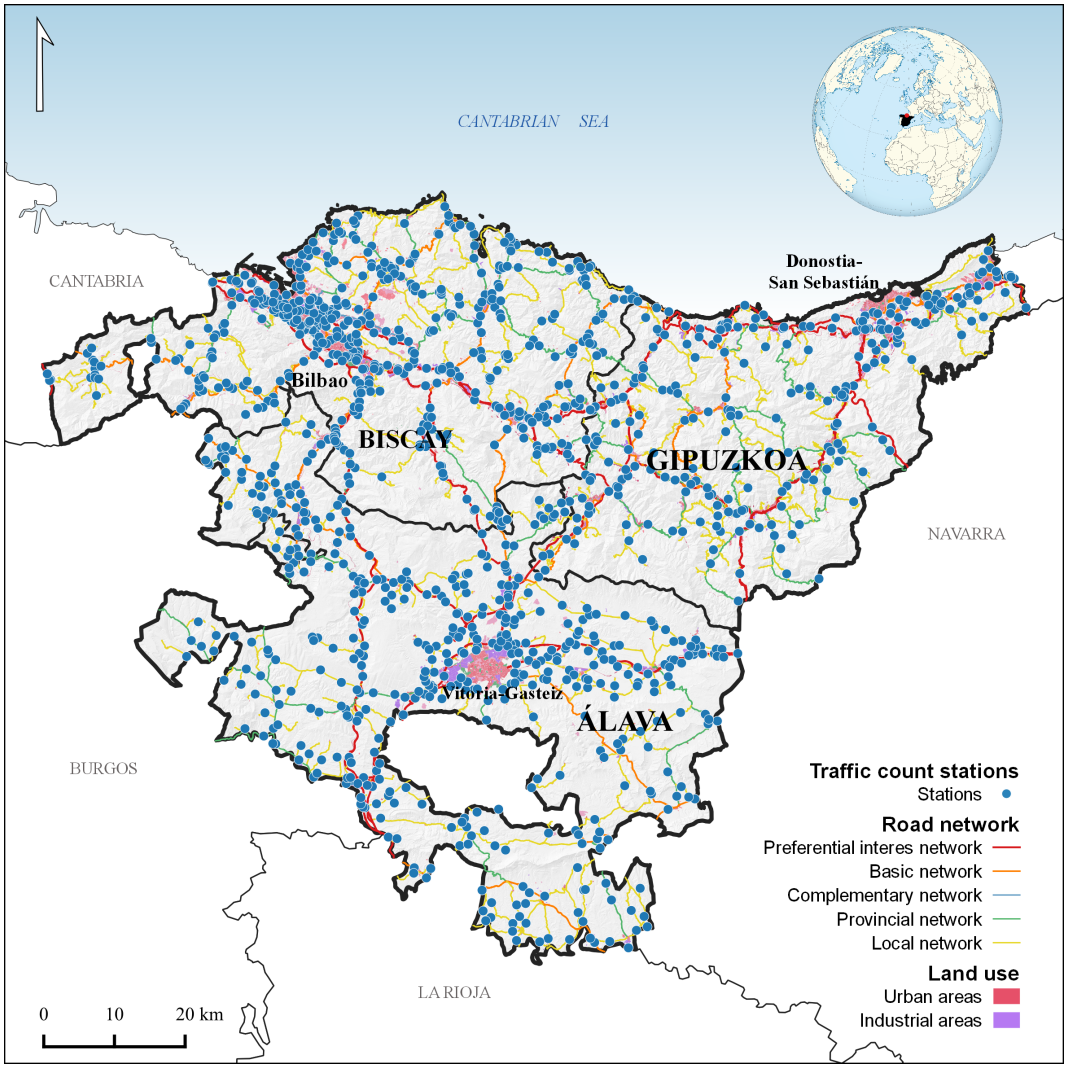

Figure 2. Location of the count stations and road network of the Basque Autonomous Community (BAC).

Table 1. Length (km) of the road network levels in each of the provinces of the BAC.

Road Network Level	Biscay	Gipuzkoa	Álava
Preferential interest network (red)	246.2	287.92	145.7
Basic network (orange)	209.1	125.98	146.42
Complementary network (blue)	32.5	-	-
Provincial network (green)	209.6	295.25	200.91
Local network (yellow)	585.9	617.07	534.23
Neighborhood network (grey)	-	-	373.02
Total	1283.3	1326.22	1400.28

In this work, four variables are considered:

- The AADT: Total number of vehicles passing a cross section of the road during a year divided by 365 (number of days of the year).
- Average annual daily heavy traffic (AADHT): Total number of heavy vehicles passing a cross section of the road during a year divided by 365 (number of days of the year).
- Number of lanes: Number of lanes in each direction.
- Speed limit: Maximum speed allowed in a specific location measured in kilometers per hour (km/h).

Initially, we had information referred to 1234 traffic count stations, from which: 511 traffic count stations are located in Biscay, 248 in Gipuzkoa, and 475 in Álava (Figure 2). Nevertheless, there were some control stations that lack information about one (or some) of the variables. The stations lacking a measurement of the variables have not been considered to do the clustering analysis. Thus, the resulting information sample was reduced by 14.99%, obtaining a statistical summary as shown in Table 2. All the stations with uncompleted data are located in Gipuzkoa, and, in the majority of the cases, they lack the average annual daily heavy traffic. Thus, information of 1049 count stations has been used in the clustering, (from which 497 count stations correspond to Biscay, 76 to Gipuzkoa, and 476 to Álava). As the AADT value was available for all te 1234 count stations, the estimation model has been developed taking into account the whole sample.

Table 2. Descriptive statistics of traffic variables.

Variable	Count	Mean	SD	Min.	Q1	Q2	Q3	Max.
AADT	1049	8578.72	16,765.34	17.00	340.00	1779.00	9174.00	140,702.00
AADHT	1049	804.14	1717.19	1.00	18.00	127.00	625.00	11,722.00
Number of lanes	1049	1.25	0.59	1.00	1.00	1.00	1.00	4.00
Speed limit	1049	60.20	20.86	30.00	50.00	50.00	70.00	120.00

2.2. Theoretical Framework

2.2.1. Clustering

In the first stage, a statistical description of the data is performed. The correlation of the main variable of interest (the AADT) and the complementary variables (the AADHT, the number of lanes, and the speed limit) is also analyzed.

In the second stage, an unsupervised machine learning algorithm is applied, concretely the k-means algorithm [41]. Unsupervised machine learning algorithms are characterized by not training previously but working directly on the data. K-means is an automatic clustering algorithm, and it is used over continuous data in exploratory levels of data analysis [42]. This algorithm is multidimensional, and its scalability is excellent. Basically, it works by grouping the data by a measure of similarity, and it requires to set the number of groups (k = 1, 2, 3, ..., n). The algorithm groups data trying to separate observations into n groups of equal variance, minimizing a criterion known as inertia.

Afterwards, the estimation model for the AADT is created using linear-type radial basis functions (RBF), generating the map of the variable. The generated model is evaluated using the leave-one-out cross validation technique.

2.2.2. Radial Basis Functions

Geostatistics studies variables distributed in an structured manner in the space, with some degree of spatial self-correlattion [43]. Let be $D \subset \mathbb{R}^d$ a domain in the space, with d being the dimension of the space ($d = 2$ for 2D and $d = 3$ for 3D), and let us consider a variable Z. This variable is regionalized if it makes correspond a value of the variable $Z(x)$ to each point x that belongs to the domain D.

Given a regionalized variable $Z(x)$, and assuming that we have n observations $Z(x_1)$, $Z(x_2)$, ..., $Z(x_n)$ of this variable, the interpolation technique called kriging consists of

predicting $Z^*(x_0)$, where $x_0 \notin \{x_1, x_2, \ldots, x_n\}$. There exist different variants of kriging (simple kriging, ordinary kriging, universal kriging, etc.), and they are given by the following expression:

$$Z^*(x_0) - m(x_0) = \sum_{i=1}^{n} \lambda_i(x_0) \cdot (Z(x_i) - m(x_i)), \quad (1)$$

with λ_i being the weight values, and $m(x_0)$ and $m(x_i)$ are the values expected by $Z(x_0)$ and $Z(x_i)$, respectively.

Radial basis function (RBF) interpolants are adequate when grid data are unavailable [44,45], and they are used when the estimation calculated by kriging is not accurate [46]. Nevertheless, in some cases, the kriging and RBF approaches are equivalent [47–49]. If we have a regionalized variable $Z(x)$ and n observations of this variable, $Z(x_1)$, $Z(x_2)$, …, $Z(x_n)$, when using RBF interpolants a function s_f is defined as follows [50,51]:

$$s_f(x) = \sum_{i=1}^{n} \alpha_i \phi(\|x - x_i\|), \quad (2)$$

where $s_f(x_k) = Z(x_k)$ is satisfied for $1 \leq k \leq n$. In our case, $\|\cdot\|$ will be the Euclidean norm in $\mathbb{R}^d$, being d the dimension of the space. The coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ are calculated by solving the linear system that comes from:

$$Z(x_k) = \sum_{i=1}^{n} \alpha_i \phi(\|x_k - x_i\|), \; 1 \leq k \leq n. \quad (3)$$

In this work, the choice for $\phi(r)$ was the linear-type RBF $\phi(r) = r$.

3. Exploratory Data Analysis

The variables AADT and AADHT are continuous and high-dispersion variables as it is shown in Table 2, and a large amount of measurements escaped from the mean in both variables. One may think the observations that escaped from the mean could be outliers. Nevertheless, they are values measured in the northern sector of the BAC (specifically in the provinces of Biscay and Gipuzkoa) where the AADT is higher. This occurs because the AP-8 freeway and the N-634 highway that connect the two major cities of the region, Bilbao and Donostia/San Sebastián, are located in the north of the region, on the coast. Additionally, the urban freeways around the metropolitan area of these two cities also concentrate high traffic volumes around them. Hence, taking into account the spatial positions that the highest values of the variables show, the decision to keep the entire sample was made, not eliminating or transforming those values called outliers.

In Figure 3, all the measurements of the four variables in each station are visualized. As it has been indicated before, the highest values of the variables AADT and AADHT are concentrated in the northern region of the BAC in the East-West direction. The same happens with the categorical variable number of lanes. Once again, as higher volumes are registered in the roads connecting the major cities (Bilbao and Donostia/San Sebastián) and in their metropolitan area, more lanes are required in these areas. However, the speed limit variable is more homogeneous. This variable does not have peaks in the northern area because the AP-8 freeway connecting Bilbao and Donostia/San Sebastián have many segments with controlled speed, under 100 km/h and even under 80 km/h, due to the geometric design in this mountainous region. Similarly, urban freeways in the metropolitan areas of these two major cities have a maximum speed of 80 km/h in most of the segments.

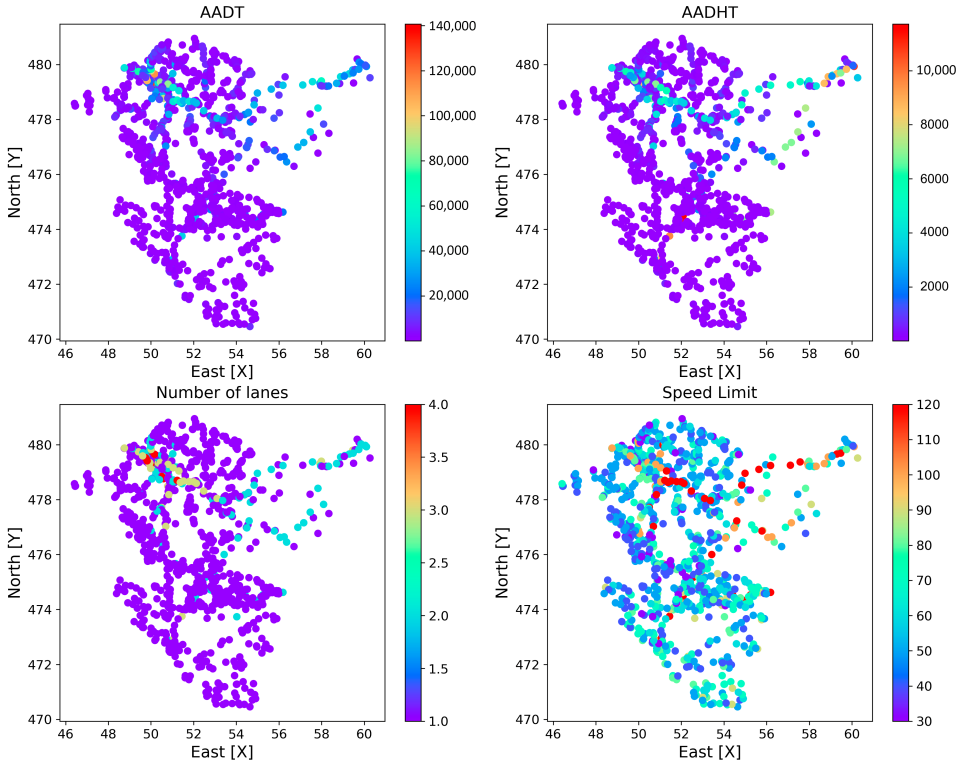

Figure 3. Values of the variables of interest per traffic count station (in 10,000 both axes).

In order to obtain a better estimate of the AADT, information of the other three variables is used, with the aim of defining domains of estimation, which are zones where the variables have a stationary behavior, and they consist in areas in which the variables result in approximately homogeneous distributions. Stationarity is a property of the model of the random function [52].

In exploratory data analysis, the existence of several populations with significantly different statistics can be indicated. The understanding of the statistical characteristics of the data leads to subdivide the area into estimation domains. The definition of estimation domains depends on the availability of sufficient data to reliably infer the statistical parameters within each of them. Furthermore, the domains must have a spatial sense and some spatial predictability, and they do not have to be excessively mixed with other domains [53]. Therefore, it is important to define whether the complementary variables really provide information for a multidimensional grouping process.

Based on the scatter diagrams and histograms of Figure 4, it can be concluded that:

- The AADT has a considerable positive correlation with the AADHT. This means that, if the circulation of heavy vehicles increases, the traffic intensity also increases. Both variables are continuous.
- The number of lanes, which is a categorical variable, also maintains a high positive correlation with the average daily intensity. Obviously, when a higher traffic volume exists, more lanes are needed.
- As regards the speed limit, this variable has a weak positive correlation with the AADT. This is understandable given that, with more traffic, it gets more difficult to travel at high speeds.

Hence, the three complementary variables provide information to the multidimensional grouping process, given the direct relationship they maintain with the variable of interest.

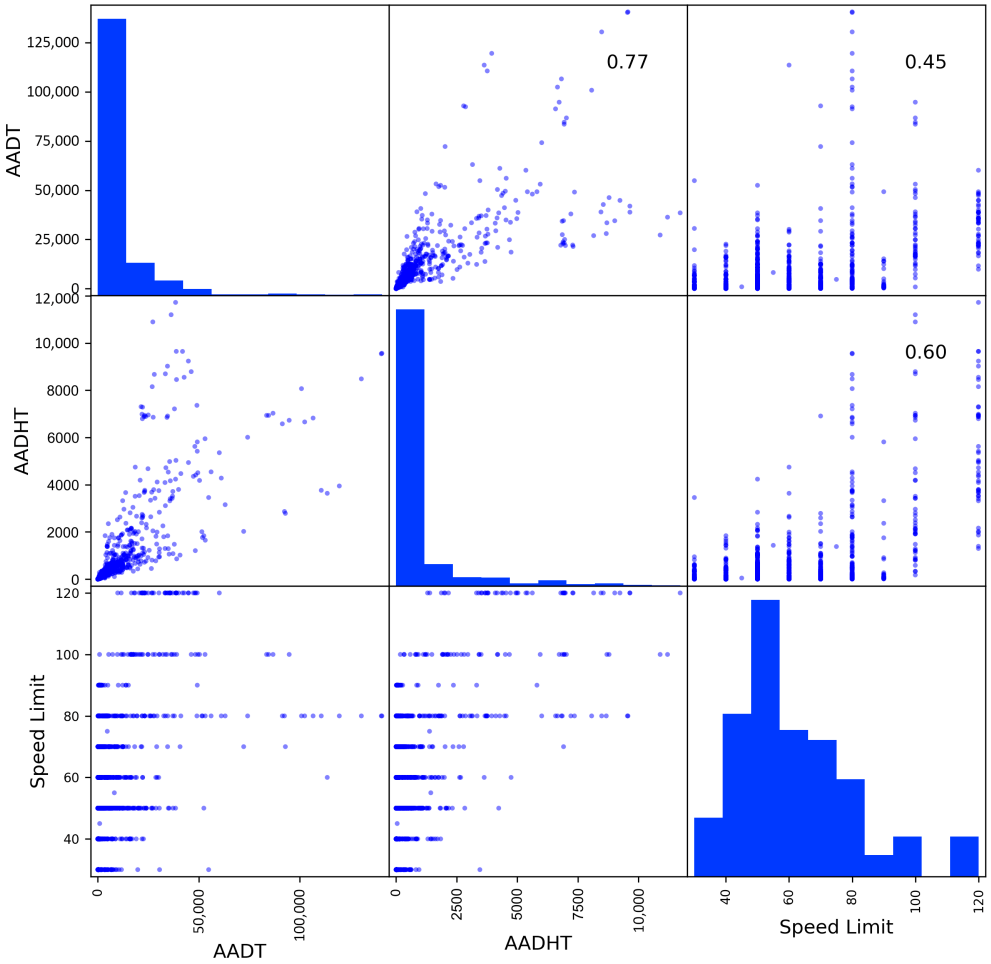

Figure 4. Scatter diagram matrix for the variables of interest of the traffic sample.

4. Clustering and Estimation Model

4.1. Clustering

The grouping is performed with the k-means algorithm based on four variables: AADT, AADHT, number of lanes, and the speed limit. The grouping process results in 3 independent groups, if the heuristic called method of the elbow is applied. Nevertheless, it has to be pointed out that the way in which this method identifies the number of groups is not always unambiguous. It is observed that, the inflection in the curve occurs with this quantity of groups, which determines the optimal number of clusters; see Figure 5. This means that the a priori estimation process should be carried out separately on 3 estimation units or clusters. In Figure 6, the positions of the count stations which conform each of the clusters are represented using different colors. And, in Table 3, the descriptive statistics of the variable of interest are shown with and without clustering.

Cluster 1 is equivalent to the 1.72% of the count stations, and it is the smallest. It shows the greatest variability as a result of grouping the highest data; it has also important peaks, and, as a consequence of this, greater differences are found. On the other hand, in clusters 2 and 3, differences of data are smaller. This fact ratifies the purpose of grouping data with the aim of homogenizing the behavior of the variable of interest.

Due to the spatial arrangement of the clusters, it may happen that estimating in each of the clusters could cause an overlap. That is to say, the data are not efficiently grouped by location; they are mixed. This is evidenced when visualizing the three clusters in Figure 7. Although the grouping hypothesis is valid and contributes in estimation processes, in this case, it is not feasible to apply it given the lack of spatial distance between clusters.

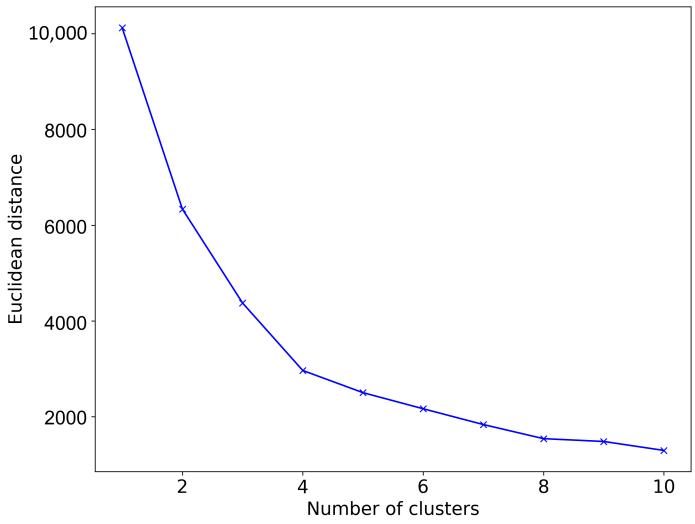

Figure 5. Elbow method showing the optimal value of clusters.

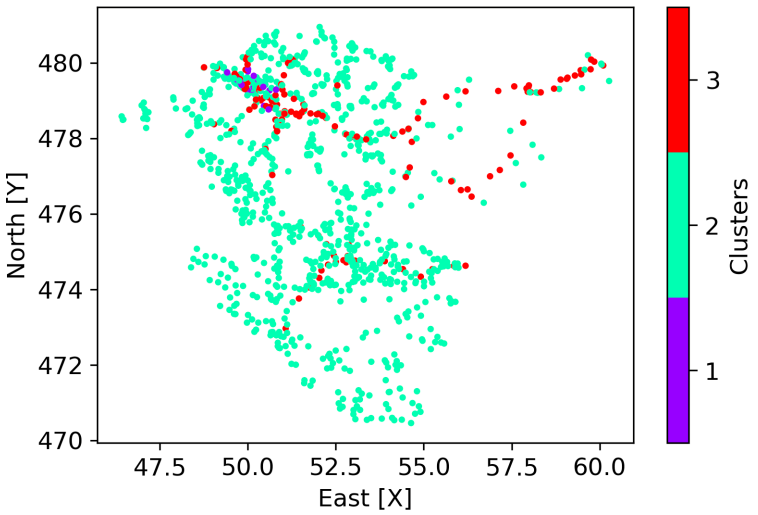

Figure 6. Spatial arrangement of the three clusters (in 10,000 m in both axes).

Table 3. Descriptive statistics of the variable average daily intensity.

AADT	Count	Mean	SD	Min.	Q1	Q2	Q3	Max.
Without grouping	1049	8578.72	16,765.34	17.00	340.00	1779.00	9174.00	140,702.00
Cluster 1	18	102,128.89	20,581.21	72,233.00	87,924.50	97,823.00	112,891.75	140,702.00
Cluster 2	891	3162.26	4186.71	17.00	261.00	1091.00	4682.50	16,997.00
Cluster 3	140	31,022.74	11,244.32	17,386.00	21,927.75	28,144.00	37,398.00	63,076.00

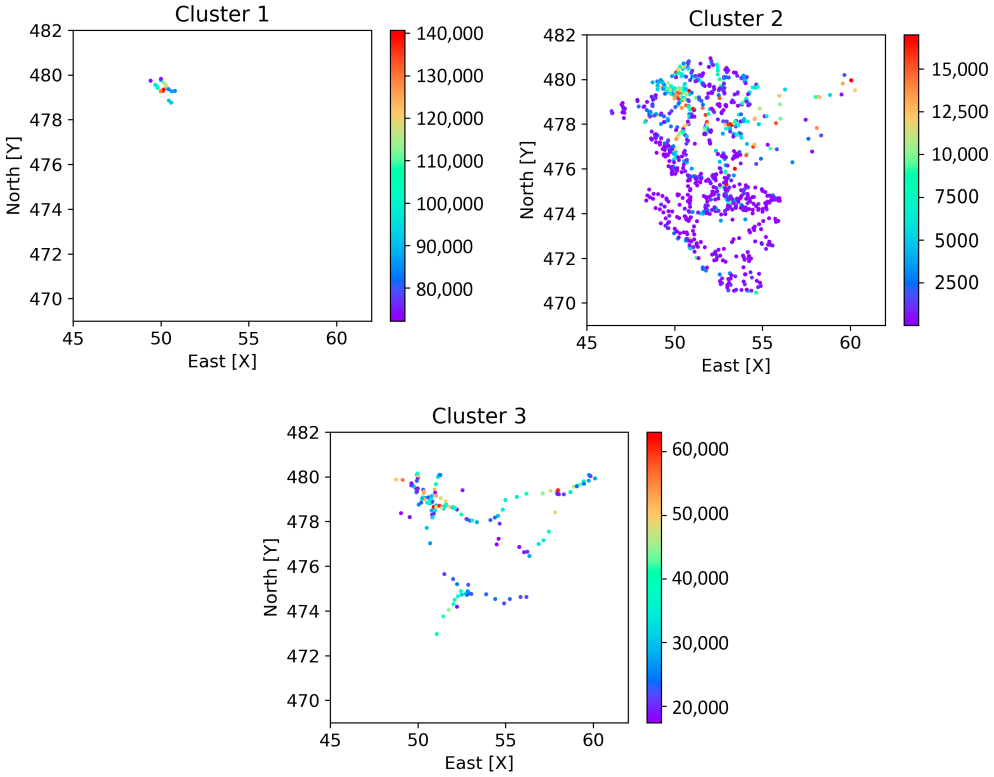

Figure 7. AADT values per count station in each cluster (in 10,000 both axes).

Cluster 1 gathers all the freeway segments around the biggest city in the Basque Autonomous Community, Bilbao, and more specifically, the metropolitan ring around it. It is normal that Bilbao, with the highest population (around 365,000 inhabitants) and its metropolitan area (with more than 1,000,000 inhabitants) concentrates the road sections with the highest values of traffic volumes, and this is reflected in the sections included in Cluster 1. Due to the high volumes, all these section have 3 or 4 lanes per direction. However, due to road safety measures and high volumes in peak hours, speed is limited to 80 km/h in some of the sections. As it could be expected, all these segments belong to the highest road network level, the red network. On the other hand, Cluster 3 groups the majority of the rest of the double carriageway roads, including freeways (tolled or not) and multilane highways and some important two-lane single carriageway roads with the highest traffic volumes. Most of the sections belong to the preferential interest network level (red network) and the basic network (the second network in importance, as shown in Table 1. However, some other roads from other network are also included, with examples of the complementary network (blue), which could be expected, but even from the provincial network (green) and the local network (yellow). This fact shows that, although the roads are classified in network according to some definitions of their function, some important quantitative variables, such as the AADT or the AADHT, reveal that they are not properly classified according to their importance. Finally, the Cluster 2 gathers the rest of the two-lane roads, and some double carriageway road segments with the lowest values in AADT. They are mainly included in the provincial, local, and neighborhood networks (green, yellow, and grey networks, respectively). Nevertheless, some sections belonging to superior networks (complementary and basic networks) can be found, even segments of the preferential interest network (red), such as segments of the tolled freeway around Bilbao with low traffic volumes.

Clustering road segments according to the AADT, AADHT, number of lanes, and speed limit helps identifying the real network levels of the roads in a territory. Sometimes, roads are included in network levels considering other criteria apart from traffic volumes, such

as the connection with adjacent territories (between the Basque Autonomous Community or with other Spanish regions, or with other countries, like France), the connection with strategic points, like ports, airports, the traditional status of some roads (they were usual routes in previous centuries, but now a better highway, generally a freeway, has modified its status), etc. These criteria, which are exposed in the laws about roads in force in each territory, do not always show the real status of a road or a road segment. Some roads can connect with adjacent provinces, but their traffic volume, both total and heavy traffic, is low because there are other connections that are preferred by drivers. Consequently, identifying the real status of a road or road segment by means of clustering helps giving the real importance to its segment, as a function of real traffic, and not based on subjective criteria. This proposed new classification can underline the real importance (or not) of each road and can help prioritizing maintenance and rehabilitation works in segments that are more important for traffic movements.

4.2. Estimation Model for the AADT

The estimation associates numerical values in those places where there is no sample. It should be noted that the estimation is at an independent level and that the complementary variables do not provide information at this stage. The estimation model has been built considering the whole sample formed by the 1234 traffic count stations and using a grid of 7236 points, see Figure 8. The descriptive statistics resulting from the model indicate a smoothing effect in the data, as it can be seen in Table 4. The variability of the AADT decreases.

Table 4. Descriptive statistics of the observed and the estimated values of the average daily intensity.

	Count	**Mean**	**SD**	**Min.**	**Q1**	**Q2**	**Q3**	**Max.**
AADT	1234	8789.20	16,912.32	17.00	386.50	1915.00	9170.25	140,702.00
AADT RBF	7236	4786.10	8841.00	0.00	227.00	1406.80	5000.30	92,661.60

The mean of the AADT of the estimated model decreases, since there are some areas where the estimation is done with poor information. This happens in the southern part of the country especially, where there are fewer roads. As a consequence, central-tendency statistics decrease, at a general descriptive statistical level.

With regard to the variability or dispersion of the data, this decreases, and this effect can be seen in the standard deviation and in the range of the values (the difference between the maximum and the minimum values). The upper bound of the estimates is significantly higher than the one of the model. The reason why the maximum value changes is because the variable is estimated on a grid that differs in position from the sample data. In other words, the created model does not estimate in the same locations of the sample, a model represented by a square mesh of 1000 × 1000 m is used and estimated from the sample.

The northeast sector of the BAC is the one in which the largest area of vehicular mobility accumulates. In particular, the highest values are concentrated in the Bilbao metropolitan area and in the surroundings of Donostia/San Sebastian. The southeast sector (which corresponds to Álava) presents a smaller vehicular area but higher average daily intensities. The southern sector has a very low level of vehicular mobility, in general.

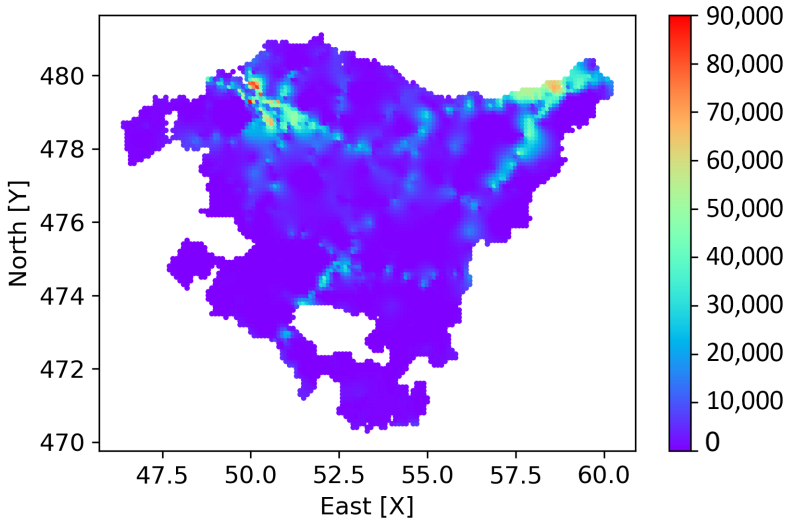

Figure 8. Annual Average Daily Traffic map generated by linear-type radial basis function (RBF) (in 10,000 m in both axes).

Model Validation

The model validation was conducted using the leave-one-out cross validation. This technique consists in choosing a location and removing the value of the selected position from the traffic sample, which is composed of $n = 1234$. Then the value is estimated using the built model. Finally, the estimated value was compared with the observed one n times. Thus, observations are divided in a training set and a single test section is performed, as many times as observations there are.

Due to the large size of the sample, we have performed a variation of the leave-one-out cross validation technique using a sub-sample of the total data. That is to say:

- Randomly, a sub-sample of 270 data of the total (equivalent to the 21.88% of the total sample ($n = 1234$)) was extracted.
- In each of the extracted points, the estimation by the RBF model is calculated using the rest of the $n - 1 = 1233$ points.
- Each of the estimated values is compared with the observed value (the real value).
- The last two steps are carried out as many times as measurements the sub-sample has (that is, 21.88% of the total sample).
- Fifty executions have been performed. The linear correlation coefficient r of each execution has been calculated, which is used as an indicator of the precision of the estimation method. The obtained best value has been $r = 0.62$, the worst $r = 0.34$, and the mean $r = 0.53$. The results that correspond to $r = 0.60$ are plotted on a scatter diagram; see Figure 9.

A first indicator of the precision of the estimate is the linear correlation coefficient. As aforementioned, the mean of 50 executions has been $r = 0.53$. This reflects a positive proportionality, but in a low range. Some measurements of the sample are separated by hundreds of kilometers from each other. Thus, when taking a measurement and estimating at that location does not produce good results. Nevertheless, interpolation methods based on linear-type radial basis functions (RBF) can be regarded as a simple and approximate technique to be applied in preliminary or early stages of the planning. Additionally, the estimate results accurate (high linear correlation coefficient) when the sub-sample is drawn from an area in which there are several close observations.

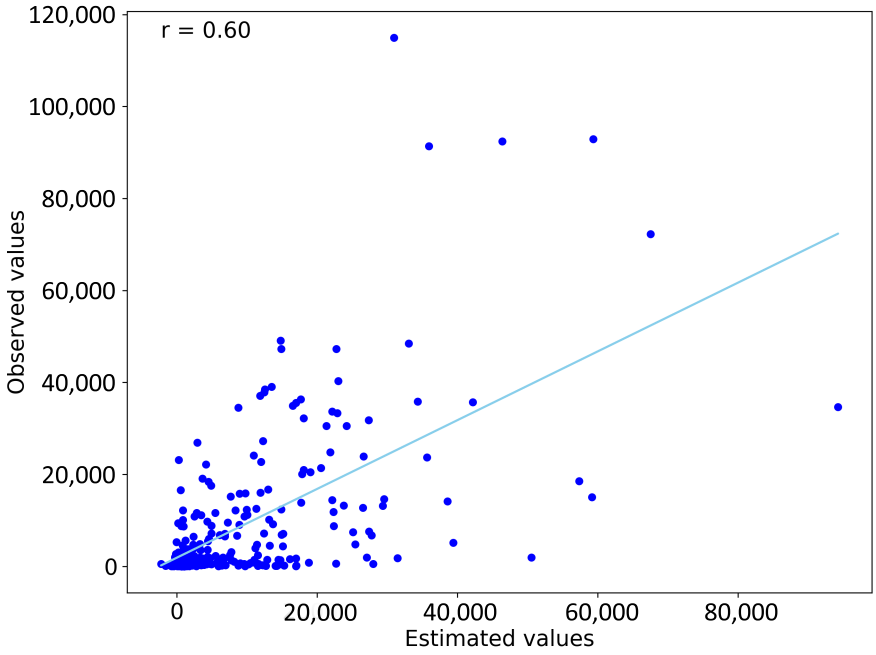

Figure 9. Cross validation of the linear-type RBF model of a execution being $r = 0.60$.

5. Conclusions

The traffic volume data is very useful information for the management of road infrastructures. More specifically, the annual average daily traffic becomes the most employed and useful variable around the world. It is used for geometric design of highways, for pavement design, and maintenance and rehabilitation strategies optimization. Therefore, knowing the exact value in each road segment is essential for any highway administration.

Clustering analysis was conducted using data for the count stations of the three provinces of the Basque Autonomous Community (BAC). The annual average daily traffic, the annual average daily heavy traffic, the number of lanes, and the speed limit were considered. Three clusters were obtained. This grouping technique is useful for identifying the real importance of each road segment, as a function of the real traffic volume it has and not based on other criteria, which cannot show the actual relevance of the section. This grouping could be useful also for the estimation process if stations were efficiently grouped spatially, which does not happen in this case.

The data about the volume of traffic intensity is, therefore, very valuable information for the administration of road infrastructures. However, obtaining accurate information for the entire network is very difficult. For that reason, the application of prediction techniques can be very helpful to improve the data obtained by the traffic count stations. Geostatistics techniques as kriging interpolation methods are the most appropriate ones to perform these analyses. This study also shows that interpolation methods based on linear-type radial basis functions (RBF) can be used as a preliminary method to estimate the annual average daily traffic. Nevertheless, in points separated by hundreds of kilometers from the count stations, the results are not very accurate. This analysis can be improved in subsequent works grouping the variable of interest into small estimation domains and making directional estimates reflecting anisotropies afterwards.

Author Contributions: All authors contributed similarly to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The University of the Basque Country (UPV/EHU), Call for Innovation Projects "IKD i^3 Laborategia" (Call 1-2020, 2019/20).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Exclude this statement.

Acknowledgments: The research described in the present paper was developed within the project entitled "Towards a Sustainable Civil Sector i3 (SSi3-CIV)" and it is based on a preliminary deliverable of this project. This project is funded by the University of the Basque Country, Project No. i320-18. The authors wish to acknowledge with thanks the Vicerectorate for Innovation, Social Commitment and Cultural Action of the University of the Basque Country (UPV/EHU) for the opportunity granted that has made possible the development of the present work.

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

References

1. EUROSTAT. *Energy, Transport and Environment Statistics—2019 Edition*; Publications Office of the European Union: Luxembourg, 2019.
2. Kraemer, C.; Pardillo, J.M.; Rocci, S.; Romana, M.G.; Sánchez Blanco, V.; del Val, M.A. *Ingeniería de Carreteras*; McGraw Hill: Madrid, Spain, 2009; Volume I.
3. Findley, D.J. Introduction. In *Highway Engineering. Planning, Design and Operations*; Findley, D.J., Schroeder, B., Cunningham, C., Brown, T., Eds.; Butterworth-Heinemann: Waltham, MA, USA, 2016; pp. 1–16.
4. Pérez-Acebo, H.; Romo-Martín, A. Service and rest areas in toll motorways in Poland: Study of distribution and facilities. *Transp. Probl.* **2019**, *14*, 155–164. [CrossRef]
5. Schroeder, B.J. Transportation Planning. In *Highway Engineering. Planning, Design and Operations*; Findley, D.J., Schroeder, B., Cunningham, C., Brown, T., Eds.; Butterworth-Heinemann: Waltham, MA, USA, 2016; pp. 17–88.
6. Monney, M.G.; Badoe, D.A.; Lee, D.J. Alternative methods for estimating seasonal factors and accuracy of daily volumes they yield. *J. Transp. Eng. Part A Syst.* **2020**, *146*, 040200007. [CrossRef]
7. Elvik, R.; Sagberg, F.; Langeland, P.A. An analysis of factors influencing accidents on road bridges in Norway. *Accid. Anal. Prev.* **2019**, *129*, 1–6. [CrossRef] [PubMed]
8. Haghani, M.; Jalalkamali, R.; Berangi, M. Assigning crashes to road segments in developing countries. *Proc. Inst. Civil Eng. Transp.* **2019**, *172*, 299–307. [CrossRef]
9. Haleem, K.; Azam, S.; Manepalli, U.; Mays, M. Identifying and comparing the injury severity risk factors on rural freeways in different states in the United States. *Int. J. Inj. Control Saf. Promot.* **2019**, *26*, 343–353. [CrossRef] [PubMed]
10. Amiri, A.M.; Sadri, A.; Nadimi, N.; Shams, M. A comparison between Artificial Neural Network and Hybrid Intelligent Genetic Algorithm in predicting the severity of fixed object crashes among elderly drivers. *Accid. Anal. Prev.* **2020**, *138*, 105468. [CrossRef] [PubMed]
11. Sun, Z.; Liu, S.; Li, D.; Tang, B.; Fang, S. Crash analysis of mountainous freeways with high bridge and tunnel ratios using road scenario-based discretization. *PLoS ONE* **2020**, *15*, e0237408. [CrossRef]
12. Chen, S.; Chen, Y.; Xing, Y. Comparison and analysis of crash frequency and rate in cross-river tunnels using random-parameter models. *J. Transp. Saf. Secur.* **2020**, 1–25. [CrossRef]
13. Paterson, W.D.O. *Road Deterioration and Maintenance Effects: Models for Planning and Management*; John Hopkins University Press: Baltimore, MD, USA, 1987.
14. Dalla Rosa, F.; Liu, L.T.; Gharaibeh, N.G. IRI Prediction Model for Use in Network-Level Pavement Management Systems. *J. Transp. Eng. Part B Pavements* **2017**, *143*, 04017001. [CrossRef]
15. Hossain, M.I.; Gopisetti, L.S.P.; Miah, M.S. International Roughness Index prediction of flexible pavements using neural networks. *J. Transp. Eng. Part B Pavements* **2019**, *145*, 04018058. [CrossRef]
16. Pérez-Acebo, H.; Mindra, N.; Railean, A.; Rojí, E. Rigid pavement performance models by means of Markov Chains with half-year step time. *Int. J. Pavement Eng.* **2019**, *20*, 830–843. [CrossRef]
17. Alaswadko, N.; Hassan, R.; Meyer, D.; Mohammed, B. Modelling roughness progression of sealed granular pavements: A new approach. *Int. J. Pavement Eng.* **2019**, *20*, 222–232. [CrossRef]
18. Zeiada, W.; Hamad, K.; Omar, M.; Underwood, B.S.; Khalil, M.A.; Karzad, A.S. Investigation and modelling of asphalt pavement performance in cold regions. *Int. J. Pavement Eng.* **2019**, *20*, 986–997. [CrossRef]
19. Pérez-Acebo, H.; Linares-Unamunzaga, A.; Rojí, E.; Gonzalo-Orden, H. IRI performance models for flexible pavements in two-lane roads until first maintenance and/or rehabilitation work. *Coatings* **2020**, *10*, 97. [CrossRef]
20. Abdelaziz, N.; Abd El-Hakim, R.T.; El-Badawy, S.M.; Afify, H.A. International Roughness Index prediction model for flexible pavement. *Int. J. Pavement Eng.* **2020**, *21*, 88–99. [CrossRef]
21. Szatkowski, W.S.; Hosking, J.R. *The Effect of Traffic and Aggregate on the Skidding Resistance of Bituminous Surfacing*; TRRL Report LR 504; Transport and Road Research Laboratory: Crowthorne, UK, 1972.

22. Kennedy, C.; Young, A.; Butler, I. Measurement of skidding resistance and surface texture and use of results in the United Kingdom. In *Suface Characteristics of Roadways: International Research and Technologies*; Meyer, W., Reichert, J., Eds.; ASTM International: West Conshohocken, PA, USA, 1990; pp. 87–102. [CrossRef]
23. Transit New Zealand (TNZ). *T10:2002. Specifications for Skid Resistance Investigation and Treatment Selection*; Transit New Zealand: Wellington, New Zealand, 2002.
24. Pérez-Acebo, H.; Gonzalo-Orden, H.; Rojí, E. Skid resistance prediction for new two-lane roads. *Proc. Inst. Civ. Eng. Transp.* **2019**, *142*, 264–273. [CrossRef]
25. Pérez-Acebo, H.; Gonzalo-Orden, H.; Findley, D.J.; Rojí, E. A skid resistance prediction model for an entire road network. *Constr. Build. Mater.* **2020**, *262*, 120041. [CrossRef]
26. Puliafito, S.E.; Allende, D.; Pinto, S.; Castesana, P. High resolution inventory of GHG emissions of road transport sector in Argentina. *Atmos. Environ.* **2015**, *101*, 303–311. [CrossRef]
27. Morley, D.W.; Gulliver, J. Method to improve traffic flow and noise exposure estimation on minor roads. *Environ. Pollut.* **2016**, *216*, 746–754. [CrossRef]
28. Liu, S.V.; Che, F.L.; Xue, J. Evaluation of traffic density parameters as an indicator of vehicle emission-related near-road air pollution: A case study with NEXUS measurement data on black carbon. *Int. J. Environ. Res. Public Health* **2017**, *14*, 1581. [CrossRef]
29. Kim, S.H.; Choi, J.G.; Ham, S.M.; Hero, W.H. Reliability evaluation of a PSC highway bridge based on resistance capacity degradation due to a corrosive environment. *Appl. Sci.* **2016**, *6*, 423. [CrossRef]
30. Zhao, F.; Chung, S. Contributing factors of annual average daily traffic in a Florida county: Exploration with geographic information system and regression models. *Transp. Res. Rec.* **2001**, *1769*, 113–122. [CrossRef]
31. Sfyridis A.; Agnolucci, P. Annual average daily traffic estimation in England and Wales: An application of clustering and regression modelling. *J. Transp. Geogr.* **2020**, *83*, 102658. [CrossRef]
32. Wu, J.Q.; Xu, H. Annual Average Daily Traffic prediction model for minor roads at intersections. *J. Transp. Eng. Part A Syst.* **2019**, *145*, 04019041. [CrossRef]
33. Ma, X.; Luan, S.; Ding, C.; Liu, H.; Wang, Y. Spatial interpolation of missing annual average daily traffic data using copula-based model. *IEEE Intell. Transp. Syst. Mag.* **2019**, *11*, 158–170. [CrossRef]
34. Khan, Z.; Khan, S.M.; Dey, K.; Chowdhury, M. Development and evaluation of recurrent neural network-based models for hourly traffic volume and annual average daily traffic prediction. *Transp. Res. Rec.* **2019**, *2673*, 489–503. [CrossRef]
35. Tawfeek, M.H.; El-Basyouny, K. Estimating traffic volume on minor roads at rural stop-controlled intersections using deep learning. *Transp. Res. Rec.* **2019**, *2673*, 108–116. [CrossRef]
36. Chang, H.H.; Cheon, S.H. The potential use of big vehicle GPS data for estimation of annual average daily traffic for unmeasured road segments. *Transportation* **2019**, *46*, 1011–1032. [CrossRef]
37. Eustat. Municipal Population Statistics (01/01//2019). Available online: www.eustat.eus (accessed on 20 August 2020).
38. Diputación Foral de Bizkaia. *Evolución del Tráfico en las Carreteras de Bizkaia—Trafikoaren Bilakaera Bizkaiko Errepideetan 2019*; Departamento de Desarrollo Económico y Territorial: Bilbao, Spain, 2020.
39. Diputación Foral de Álava. *Estudio de Tráfico. 2019. Red de Carreteras del Territorio Histórico de Álava—2019. Trafiko Azterketa. Arabako Lurralde Historikoaren Errepide-Sarea*; Departamento de Infraestructuras Viarias y Movilidad: Vitoria-Gasteiz, Spain, 2020.
40. Diputación Foral de Gipuzkoa. *Información de Aforos en las Carreteras de Gipuzkoa. Recopilación Hasta 2019*; Departamento de Infraestructuras Viarias: Donostia-San Sebastián, Spain, 2020.
41. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Symposium of Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1965; University of California Press: Berkeley, CA, USA, 1965; Volume 1, pp. 281–297.
42. Fukunaga, K. *Introduction to Statistical Pattern Recognition*; Academic Press: San Diego, CA, USA, 1990.
43. Sinclair, A.J.; Blackwell, G.H. *Applied Mineral Inventory Estimation*; Cambridge University Press: Cambridge, UK, 2002.
44. Hardy, L.R. Multiquadric equations of topography and other irregular surfaces. *J. Geophys. Res.* **1971**, *76*, 1905–1915. [CrossRef]
45. Buhmann, M.D. Radial basis function. *Acta Numer.* **2000**, *9*, 1–38. [CrossRef]
46. Elsayed, K.; Lacor, C. Robust parameter design optimization using Kriging, RBF and RBFNN with gradient-based and evolutionary optimization techniques. *Appl. Math. Comput.* **2014**, *236*, 325–344. [CrossRef]
47. Watson, G.S. Smoothing and interpolation by kriging and with splines. *Math. Geol.* **1984**, *16*, 601–615. [CrossRef]
48. Myers, D.E. Interpolation with positive definite functions. *Sci. Terre* **1988**, *28*, 252–265.
49. Cressie, N. Geostatistics. *Am. Stat.* **1989**, *43*, 197–202.
50. Buhmann, M.D. *Radial Basis Functions-Theory and Implementations*; Cambridge University Press: Cambridge, UK, 2003.
51. Guttman, H.M. A radial basis function method for global optimization. *J. Glob. Optim.* **2001**, *19*, 201–227. [CrossRef]
52. Isaaks, E.H.; Srivastava, M.R. *Applied Geostatistics*; Oxford University Press: New York, NY, USA, 1989.
53. Coombes, J. *The Art and Science of Resource Estimation: A Practical Guide for Geologists and Engineers*; Coombes Capability: Perth, Australia, 2008.