International Journal of Geo-Information

MDPI

Article

Make It Simple: Effective Road Selection for Small-Scale Map Design Using Decision-Tree-Based Models

Izabela Karsznia [1,*], Karolina Wereszczyńska [1] and Robert Weibel [2]

[1] Department of Geoinformatics, Cartography and Remote Sensing, Faculty of Geography and Regional Studies, University of Warsaw, Krakowskie Przedmiescie 30, 00-927 Warsaw, Poland
[2] Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
* Correspondence: i.karsznia@uw.edu.pl

Abstract: The complexity of a road network must be reduced after a scale change, so that the legibility of the map can be maintained. However, deciding whether to show a particular road section on the map is a very complex process. This process, called selection, constitutes the first step in a sequence of further generalization operations and it is a prerequisite to effective road network generalization. So far, not many comprehensive solutions have been developed for effective road selection specifically at small scales as the studies have mainly dealt with large-scale maps. The paper presents an experiment using machine learning (ML), specifically decision-tree-based (DT) models, to optimize the selection of the roads from 1:250,000 to 1:500,000 and 1:1,000,000 scales. The scope of this research covers designing and verifying road selection models on the example of three selected districts in Poland. The aim is to consider the problem of road generalization holistically, including numerous semantic, geometric, topological, and statistical road characteristics. The research resulted in a list of measurable road attributes that comprehensively describe the rank of a particular road section. The outcome also includes attribute weights, attribute correlation calculated for roads, and machine learning models designed for automatic road network selection. The performance of the machine learning models is very high and ranges from 80.94% to 91.23% for the 1:500,000 target scale and 98.21% to 99.86% for the 1:1,000,000 scale.

Keywords: roads; cartographic generalization; decision trees; small-scale maps; machine learning

Citation: Karsznia, I.; Wereszczyńska, K.; Weibel, R. Make It Simple: Effective Road Selection for Small-Scale Map Design Using Decision-Tree-Based Models. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 457. https://doi.org/10.3390/ijgi11080457

Academic Editors: Florian Hruby and Wolfgang Kainz

Received: 14 June 2022
Accepted: 29 July 2022
Published: 22 August 2022

1. Introduction

Generalization of the road network has had the attention of scientists for decades [1–4]. Various methods for selecting roads at large map scales have been developed [5], but there is still a lack of effective solutions for small scales [6]. Thus, in this research we concentrate on small scales, but we also believe that, after appropriate modifications, the proposed approach can be applied at large scales.

Researchers agree that an adequate road network generalization process requires that consideration be given to numerous semantic, geometric, topological, and statistical road network characteristics [7–10]. For the road selection process, various approaches have been proposed, including graph-theory-based methods [8,11], stroke-based methods [9], enhanced central algorithm [4], methods based on information theory, and various other measures. However, graph-theory-based methods usually do not consider semantic and geometric roads characteristics, while stroke-based methods rarely take into account statistical road characteristics [9,12].

Experienced cartographers make decisions based on a multitude of conditions and dependencies, while simultaneously considering various map object characteristics. An effective and automatic algorithm holistically elaborating essential road characteristics would help to reduce the costs of map design, while making it more efficient and faster. Replacing the cartographer's subjective decisions with an algorithm requires that a number

of map object characteristics be taken into account that, for the sake of process automation, must be measurable and comparable. Developing such advanced algorithms also requires the use of tools that allow large datasets to be processed and relevant regularities to be searched for. Machine learning (ML), which is successfully used in cartography and many other domains, provides such opportunities. This approach has proven to be a promising solution for settlement selection at small scales [13–15] and generalization of buildings [16], also with the use of deep learning (DL) [17], as well as for smoothing and selecting of line objects [18,19], especially with the use of neural networks [20].

Although traditional artificial intelligence and the graph theory-based methods have improved road selection efficiency, they also have some drawbacks. For instance, they cannot fully extract spatial characteristics of road networks; moreover, they are not always automatic and frequently subjective [21]. The solution can be combining those approaches and using graph convolutional networks (GCNs), which benefit from artificial intelligence and graph theory [22]. In the study [21], deep graph convolutional networks (DGCNs) were applied for automatic road selection. The authors compared and analyzed the results of various GCNs (GraphSAGE and graph attention networks [GAT]) by selecting small-scale road networks with the use of different deep architectures (JK-Nets, ResNet, and DenseNet). They concluded that the GAT and JK-Nets architectures provided the best results, with an accuracy of around 88% in comparison to selection conducted by an expert. Nevertheless, even in the case of GAT, the authors noticed a number of both incorrectly selected and deleted roads, thus concluding that such research requires further experiments as they are still far from practical applications [21].

Another interesting approach applied in road network selection is the use of an analytical hierarchy process (AHP). AHP is one of the multicriteria decision-making (MCDM) methods. This technique helps to find optimal decisions by considering multiple values and user preferences. Conventional AHP is usually used in the case of clearly defined decision applications, which is not always the case in road selection. In [23], the authors proposed and implemented a fuzzy-AHP approach assuming road attributes as fuzzy criteria, then used this approach to define attribute weights and used them to build a road network hierarchy, to set up priorities of the roads for selection [23]. In this approach five attributes concerning semantic, geometric, and topological road properties were used, namely road class, length, and three centrality measures (betweenness, closeness, and degree). The classical AHP approach was also recently used for road selection by [24]. They employed points of interest (POIs) to build indicators of contextual characteristics and calculate particular stroke importance with the use of the AHP model. Roads were then selected based on stroke importance, as well as on further criteria of density and overall road network connectivity maintenance. Based on the resulting map at 1:200,000 scale, the authors concluded that using the AHP method helps to preserve the structure and characteristics of the source road network.

The shortcomings of the described research included considering only selected road characteristics. Rarely, extensive semantics together with statistical, geometrical, and topological road features were taken into account. In addition, the previous research requires further experiments and extensions as they are still far from practical applications. Moreover, the proposed solutions are in the majority dedicated and adapted to large-scale maps. Thus, the motivation of this research is to fill this research gap and develop fully automated methods and models, dedicated but not limited to small scales and taking into account rich and holistic road network measures.

The specific aim of this study is to evaluate and extend the method based on data enrichment and machine learning, proposed for settlements by [13]. The motivation is to enrich the method with new variables, defined for the road network, thus extending the proposed method to the road network, as well as to consider new test areas and further machine learning models. Such an extension allows us to assess the universality of the developed methodology and the possibility of its application on various cartographic objects of different geographical characteristics.

2. Materials and Methods

2.1. Data

The source data constitutes the road network contained in the General Geographic Object Database (GGOD) corresponding to a 1:250,000 scale. The GGOD covers the whole country and is obtained by semi-automatic generalization from the Topographic Objects Database. In this study, three test areas were considered, i.e., three districts in Poland. The districts vary in terms of road network characteristics (Figure 1). The areas are located in different parts of Poland and depict the diversity of the road network density for the whole country. The reference to administrative units of districts is related to their optimal size for small-scale data analyses. The existing regulation on mapping also indicates districts as basic units for the generalization of spatial data [25].

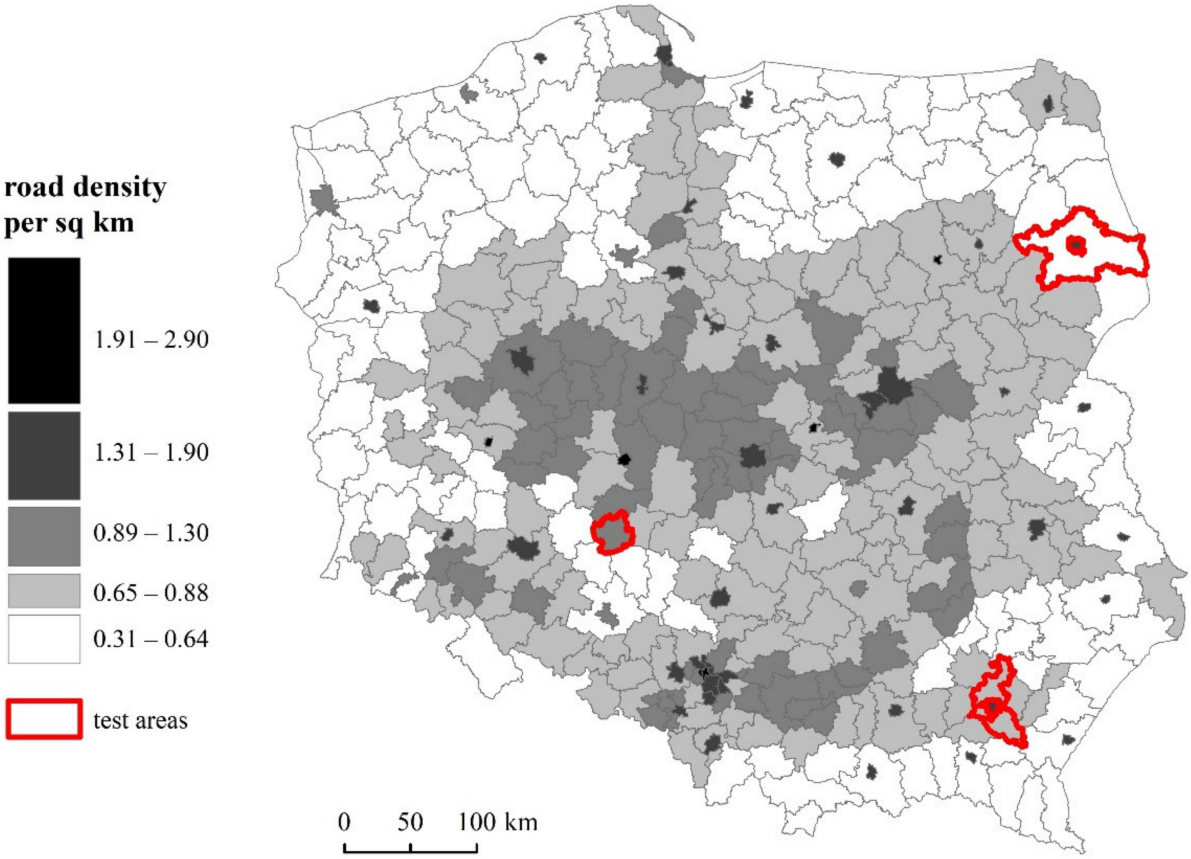

Figure 1. Districts selected as test areas overlaid on a choropleth map of road density in Poland.

2.2. Selection Based on Regulation and Machine Learning

Within the scope of this research, two approaches have been designed, namely basic and enhanced (Figure 2).

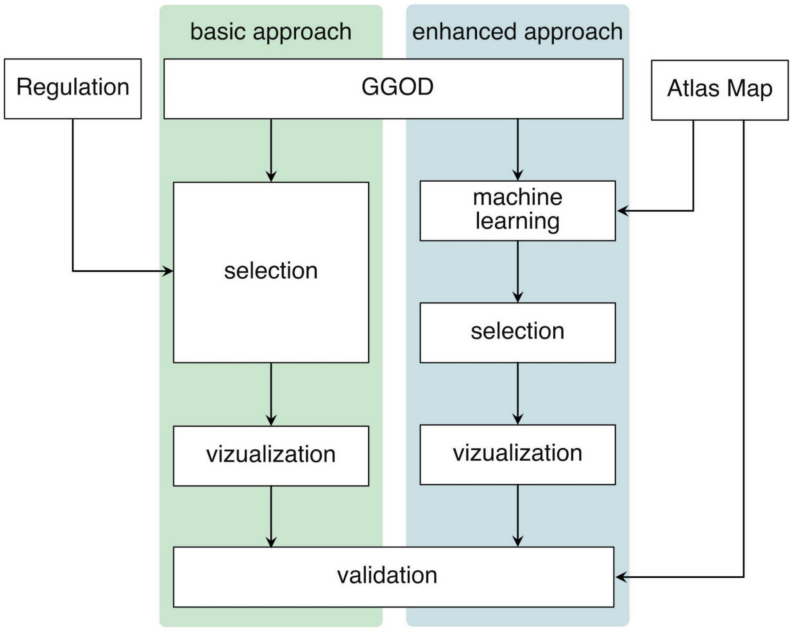

Figure 2. Workflow of research methodology (based on [13]).

In the basic approach, the selection rules were obtained from the regulations of the Polish Ministry of Interior and Administration [25] and applied to the GGOD road data. The basic approach was implemented using ArcMap version 10.6 functionality. The attributes of the roads taken into consideration in the basic approach are presented in Table 1.

Table 1. Road variables and their formalization implemented in the basic approach (based on [25]).

	Variable	Description
attribute variables	road class	hierarchy of roads in terms of technical condition, featuring: motorway, expressway, main road of accelerated traffic, main road, collector road, local road, local access road
	road category	hierarchy of roads in terms of function, featuring: national roads, voivodeship roads, district roads, municipality roads
	number of carriageways	number of road carriageways
	type of surface	type of road surface—paved or unpaved

In the enhanced approach, the following steps were conducted: (1) collecting cartographic knowledge, (2) enriching roads with additional characteristics (named variables), (3) formalizing variable values, (4) using machine learning models, namely decision trees (DT), optimizing decision trees with genetic algorithms (DT-GA) and random forest (RF), (5) implementing the developed models, and (6) validating the results.

Within the first step, experienced cartographers were consulted. Issues such as road characteristics, relations with other map objects, road network continuity, and road network patterns maintenance were discussed. In step two, the list of essential road characteristics was gathered. In this study, the characteristics measurably formalized are called variables. The variables concerned both road semantic attributes and spatial characteristics. The variables considered in the enhanced approach are presented in Table 2. The transition from general ideas of variables for roads to specific quantifiable values has been called "variable value formalization". All additional values were calculated using ArcGIS version 10.6 and Python tools. The road network consists of many sections (Figure 3). A road between intersections is called a "segment" and may consist of multiple "sections". In this research we build sections based on attribute values, namely road class and category, number of carriageways, and type of surface, while the segments may constitute several

sections. The segments were created by connecting road sections between intersections. The object called in this case a segment corresponds to the edges of the graph, and the intersections are its nodes (uniform in terms of attribute values). In step three, a method of calculating values for each variable was developed. Obtaining quantified values that can be objectively calculated and compared is most relevant for the process automation.

Table 2. Road variables and their formalization implemented in the enhanced approach.

	Variable	Description
attribute variables	road class	hierarchy of roads in terms of technical condition, featuring: motorway, expressway, main road of accelerated traffic, main road, collector road, local road, local access road
	road category	hierarchy of roads in terms of function, featuring: national roads, voivodeship roads, district roads, municipality roads
	number of carriageways	number of road carriageways
	type of surface	type of road surface—paved or unpaved
spatial variables	segment length	segment length in metric units
	no. of connected roads (segment)	number of roads linked to a road segment that the road section is a part of
	no. of connected roads (section)	number of roads linked to a road section
	connects the settlements	number of settlements the road segment connects
	minimum number of segments leading from settlements at a scale of 1:500,000	minimum number of road segments exiting from settlements selected as a result of ML (scale 1:500,000) with which the segment connects
	minimum number of segments leading from settlements at a scale of 1:1,000,000	minimum number of road segments exiting from settlements selected as a result of ML (scale 1:1,000,000) with which the segment connects
	density of paved roads in the district	density of paved road network in the district (in km/100 km^2)
	density of paved roads in hexagon	density of paved road network in the hexagon (in km/100 km^2)
	density of roads in the district	density of road network in the district (in km/100 km^2)
	density of roads in a hexagon	density of road network in the hexagon (in km/100 km^2)
	betweenness centrality	an indicator of an edge's centrality in a network

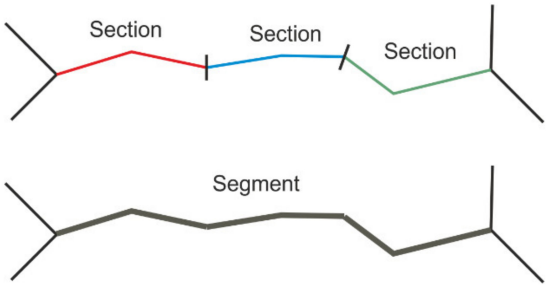

Figure 3. Roads presented as sections and segments.

We also investigated the relations with settlements and other sections of the network, such as road network density within the district, number of connections with other roads, edge betweenness centrality measure, type of surface, and number of lanes. These variables had to be measurable and comparable among the roads constituting the road network [26]. Enriching the road data with the relevant variables and considering sufficiently large samples, we were able to design the machine learning models.

In step four of the enhanced approach, ML was executed in RapidMiner Studio version 9.9 (Technical University of Dortmund, Germany). The selection process consisted of developing and applying the three decision-tree-based models. We designed separate selection processes based on decision trees (DT), decision trees supported with genetic

algorithms (DT-GA) and random forest models (RF). For all tested ML models, the default parameters implemented in the data mining software were used [27]. Parameters and their values were as follows:

- Gain ratio: In the case of the DT model the criterion on which attributes were selected for splitting was the gain ratio. It is a variant of information gain that adjusts the information gain for each attribute to allow the breadth and uniformity of the attribute values to be captured. The gain ratio parameter calculates the weight of attributes with respect to the label attribute by using the information gain ratio. The higher the weight of an attribute, the more relevant it is considered.
- Minimal split size: Set equal to 4. The size of a node is the number of examples in its subset. The size of the root node is equal to the total number of examples considered. Only those nodes are split whose size is greater than or equal to the minimal split size.
- Minimal leaf size was equal to 4. The size of a leaf node is the number of examples in its subset. The tree is generated in such a way that every leaf node subset has at least the minimal leaf size number of instances.
- Minimal gain: Set equal to 0.1. The gain of a node is calculated before splitting it. The node is split if its gain is greater than the minimal gain. Higher values of the minimal gain result in fewer splits and thus a smaller tree. A too high value will completely prevent splitting and a tree with a single node is generated.
- Minimal depth: Set equal to 20. The depth of a tree varies depending upon size and nature of the examples. This parameter is used to restrict the size of the decision tree.
- Confidence: Set equal to 0.25. This parameter specifies the confidence level used for the pessimistic error calculation of pruning.
- Number of prepruning alternatives: Set equal to 3. This parameter adjusts the number of alternative nodes tried for splitting when a split is prevented by prepruning at a certain node.
- In the case of the DT-GA model, the operator named 'optimize selection' was additionally used, invoking a genetic algorithm (GA) to select the most relevant attributes of the given sample set. A GA is a search heuristic that mimics the process of natural evolution, such as inheritance, mutation, selection, and crossover [27].
- For the RF model, further parameters included the number of trees to be generated (set to 100) and the maximal tree depth (set equal to 10). This parameter is used to restrict the depth for each random tree set [27].

Roads presented on atlas maps designed by experienced cartographers were used as the training material. As a result, we obtained decision trees that showed which variables are decisive when selecting particular road segments. Based on the model derived from the learning process, the prediction process was conducted. Then, the roads meeting the model requirements were selected. In both approaches, the correctness of selection was assessed by comparing the results of the basic and enhanced approaches against the roads shown on the atlas maps designed by experienced cartographers. As a result of the application of ML models, we also received variable weights and variable correlations.

3. Results

3.1. Decision Trees

The research results constitute the decision-tree-based models, a correlation matrix, road variable weights and the roads generalized to 1:500,000 and 1:1,000,000 scales. From the obtained decision trees, the most significant variables in the road selection process can be identified. They are placed at the root of the tree (Figures 4–6).

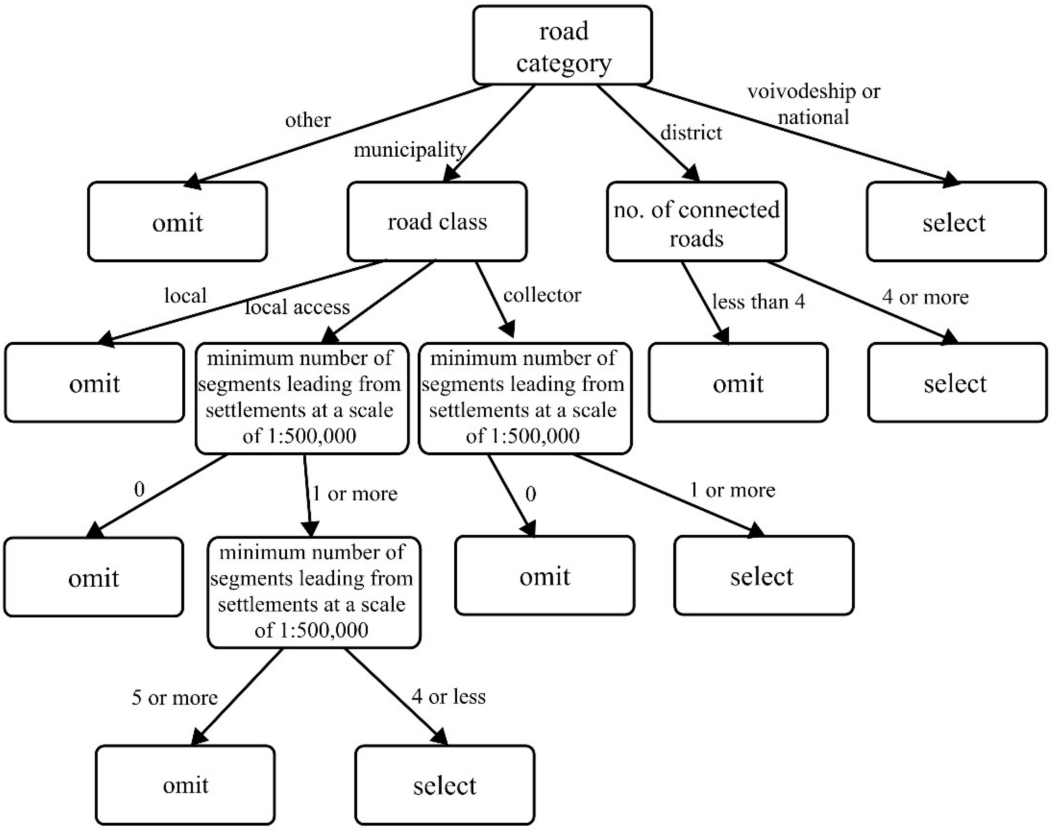

Figure 4. Decision tree for the three districts, the result of machine learning DT-GA—road selection for 1:500,000 scale.

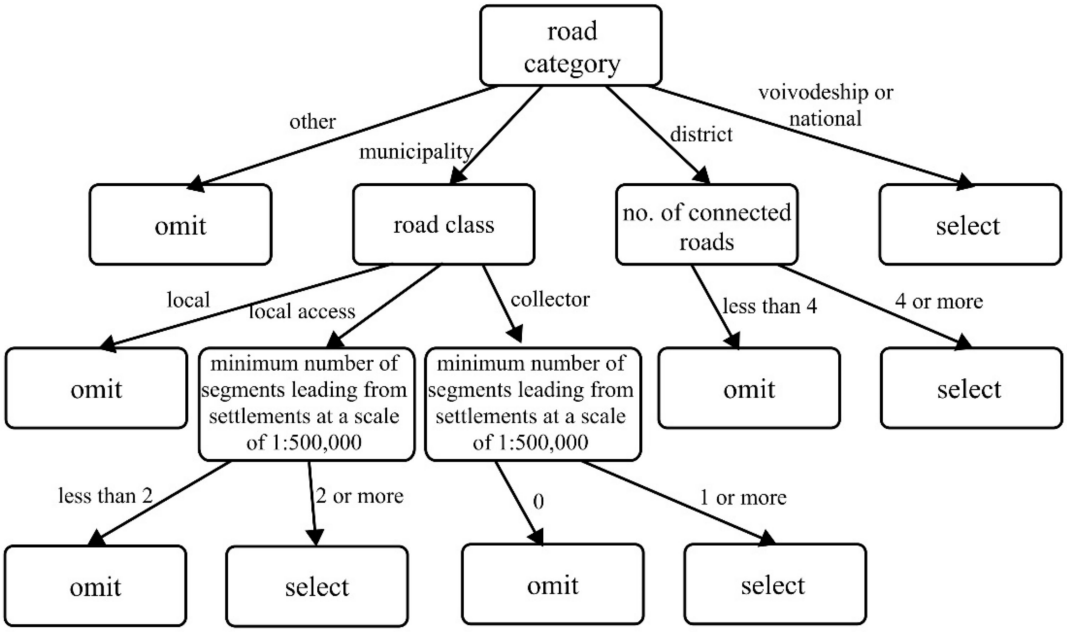

Figure 5. Decision tree for the three districts, the result of machine learning DT—road selection for 1:500,000 scale.

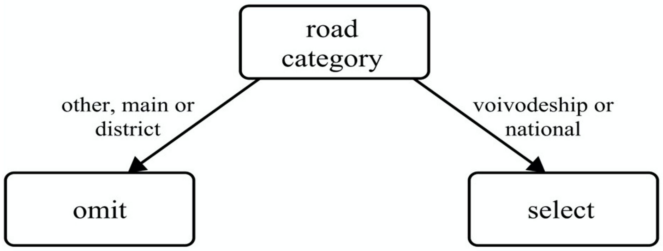

Figure 6. Decision tree for the three districts, the result of machine learning DT and DT-GA—road selection for 1:1,000,000 scale.

Thanks to the ability to read the rules from the decision tree (DT and DT-GA), it is possible to pre-evaluate the plausibility of machine learning models. The developed rules are logical, and they are also in line with generally accepted cartographic practice.

When using the RF algorithm, multiple decision trees are created, making it impossible to supervise and evaluate machine learning in this approach. The black-box nature of this method, however, constitutes a significant disadvantage.

3.2. ML Model Accuracy

The accuracy of the obtained results was determined by assessing the similarity percentage between the achieved results and the atlas map designed by a cartographer. The accuracy was calculated as the number of road segments classified as selected or omitted both on the atlas map and in either the basic or enhanced approach. The accuracy of the basic approach and all implemented ML models is presented in Tables 3 and 4, respectively, for the two considered levels of detail.

Table 3. Selection accuracy for 1:500,000 scale. Highest performance values in bold. For the difference calculation, the best performing model of the enhanced approach was used.

Area	Basic Approach	DT	DT-GA	RF	Difference
All districts	45.10%	82.46%	83.33%	**84.96%**	+39.86%
Białostocki	43.70%	74.39%	75.98%	**80.94%**	+37.24%
Rzeszowski	55.25%	84.38%	**86.76%**	82.58%	+31.51%
Kępiński	42.10%	86.55%	89.62%	**91.23%**	+49.13%

Table 4. Selection accuracy for 1:1,000,000 scale. Highest performance in bold. For the difference calculation, the best performing model of the enhanced approach was used.

Area	Basic Approach	DT	DT-GA	RF	Difference
All districts	96.36%	99.18%	**99.44%**	99.34%	+3.08%
Białostocki	96.46%	99.05%	**99.79%**	99.58%	+3.33%
Rzeszowski	91.89%	96.11%	**98.21%**	97.00%	+7.32%
Kępiński	98.39%	99.71%	**99.86%**	99.71%	+1.47%

3.3. Maps

The results of the basic and enhanced approach were also visually compared against the atlas map to assess network consistency and maintenance of characteristic road network patterns. Figures 7–12 provide the results for the considered districts for both levels of detail.

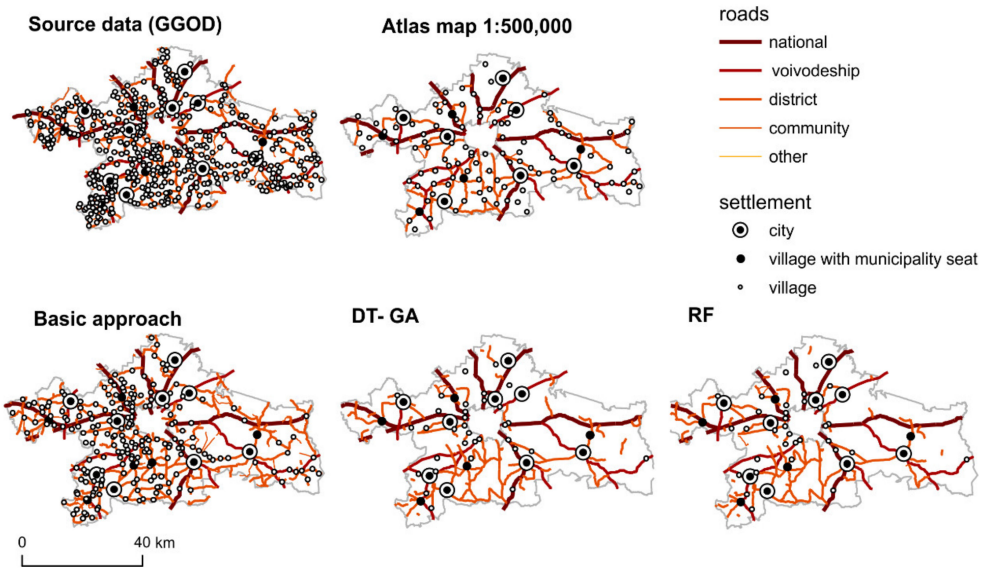

Figure 7. Selection results in Białostocki district at 1:500,000 scale.

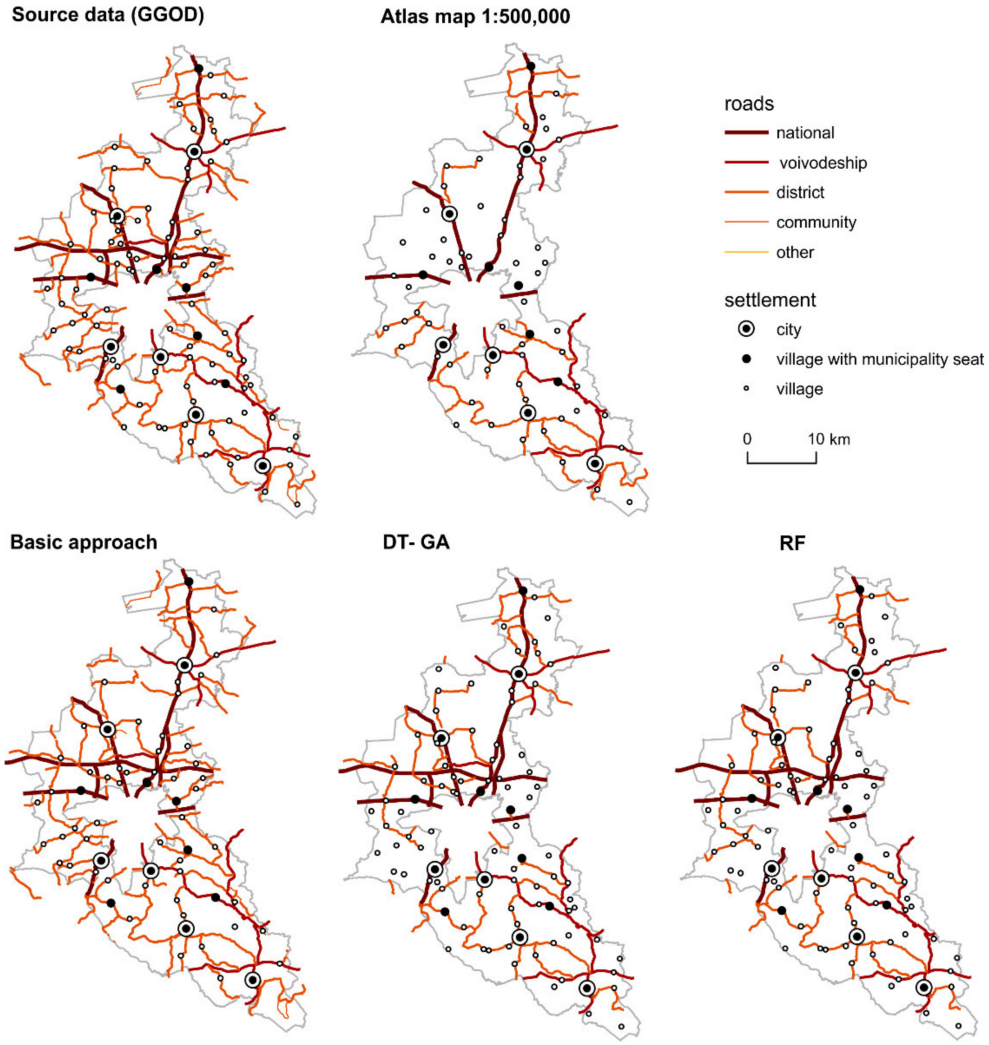

Figure 8. Selection results in Rzeszowski district at 1:500,000 scale.

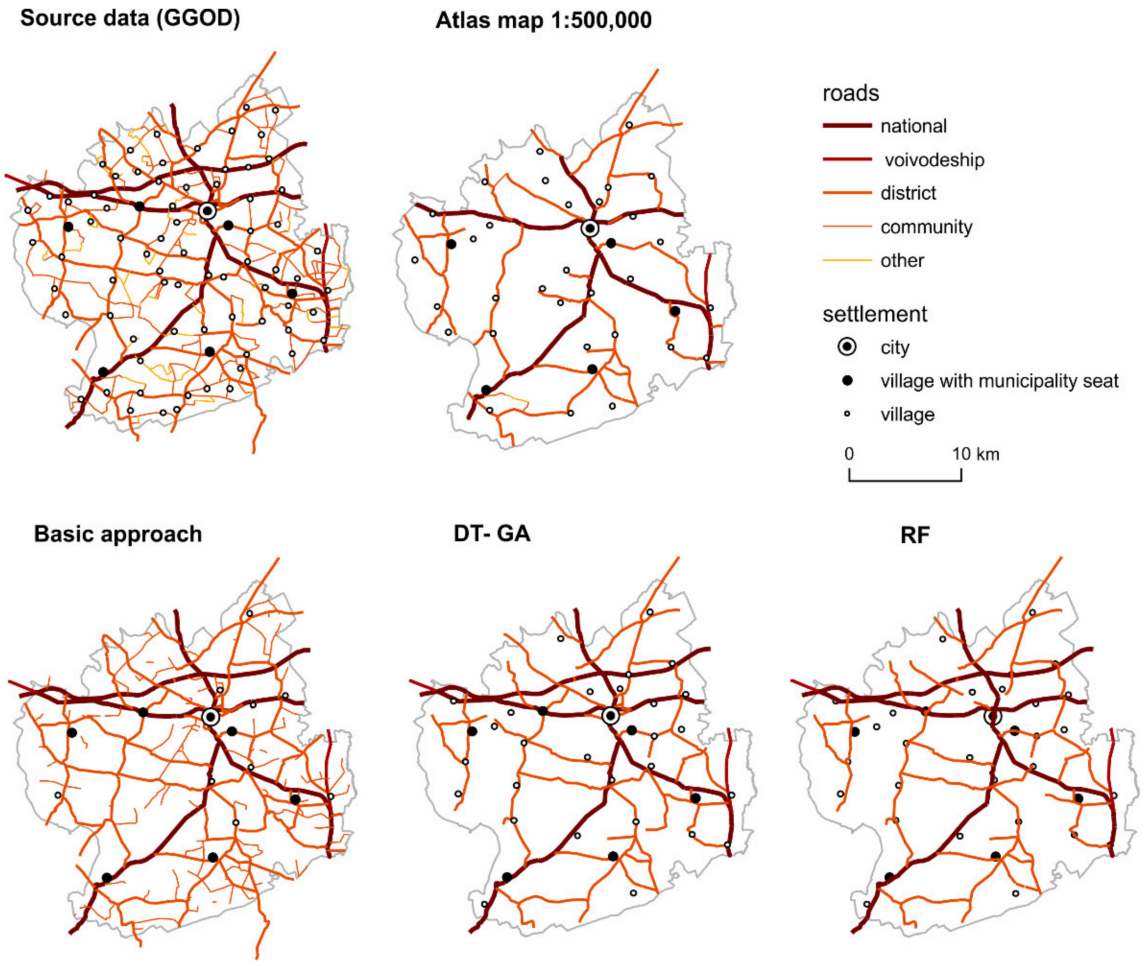

Figure 9. Selection results in Kępiński district at 1:500,000 scale.

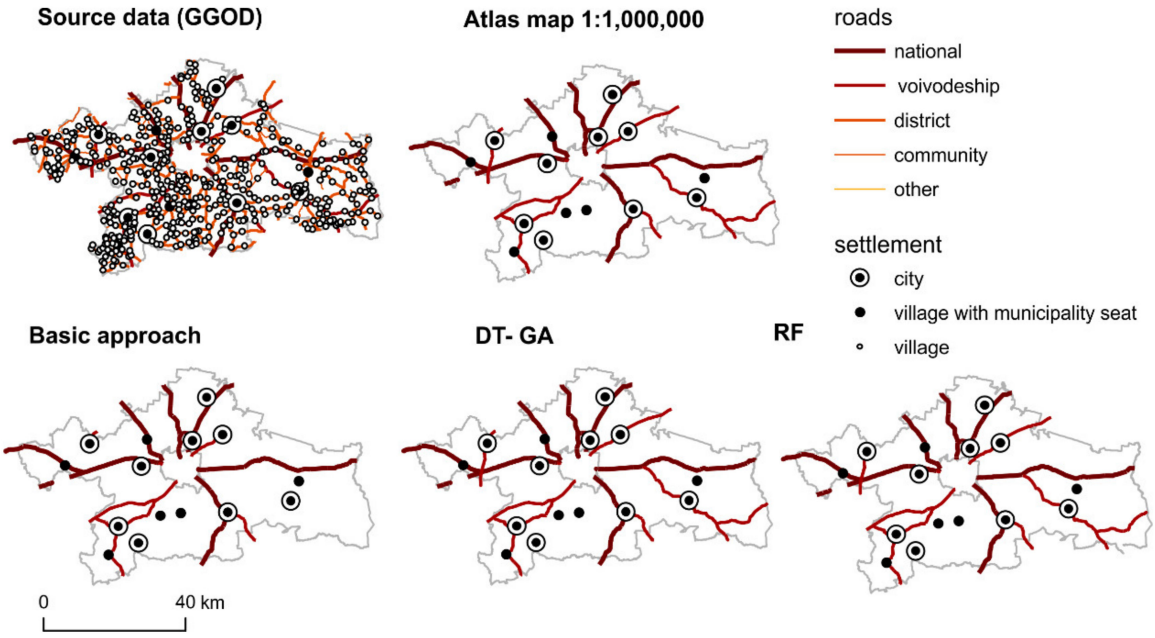

Figure 10. Selection results in Białostocki district at 1:1,000,000 scale.

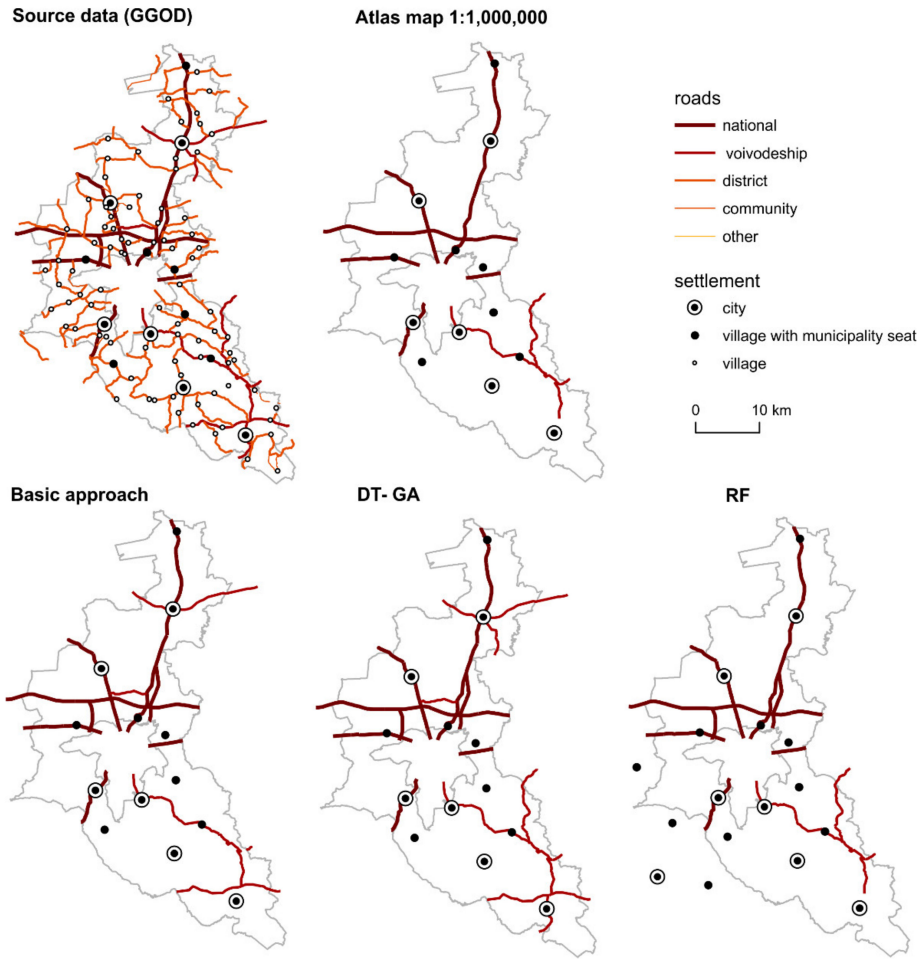

Figure 11. Selection results in Rzeszowski district at 1:1,000,000 scale.

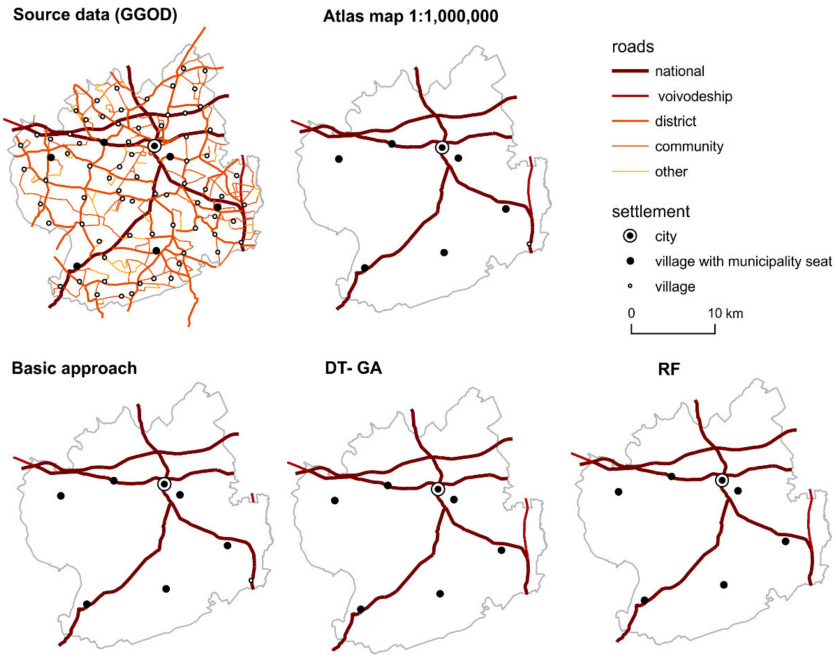

Figure 12. Selection results in Kępiński district at 1:1,000,000 scale.

4. Evaluation

To verify the suitability of the decision-tree-based models for automatic road selection in small-scale maps, we designed three models for the initial sample of three districts in Poland. We evaluated decision trees (DT), decision trees supported with genetic algorithms (DT-GA), and random forest (RF) machine learning models. The goal was to consider new variables, examine their importance and correlation, and assess the automatic machine learning models' outcomes qualitatively and quantitatively. The section is organized to comment on these goals and challenges.

4.1. Qualitative Assessment

A visual assessment of the results obtained from the enhanced and basic approaches in comparison to the atlas map favors the enhanced approach (see Figures 7–12). In the enhanced approach, the main roads have been preserved, the selected roads constitute a more coherent network in relation to the basic approach, and the general character and differences in the road density are better reflected than in the basic approach. Meanwhile, in the basic approach, the road network is too dense, as compared to the atlas map (Figures 8 and 9) and is also discontinuous in many places (Figures 7–9). At the same time, Figure 12 shows that in both enhanced approaches (DT-GA and RF), the important road in the eastern part of the district was properly kept on the map, while in the basic approach it was omitted. On the other hand, in Figure 11, some roads were omitted in the southern district part by RF and the basic approach, which makes the RF outcome, in particular, more similar to the atlas map. Meanwhile, these roads were kept in the DT-GA map, so, although it is less similar to the atlas map, this solution makes the DT-GA outcome more correct in terms of maintaining overall road network consistency in this district. Based on the qualitative assessment, we also note problems with dead-end roads in basic and enhanced approaches alike (Figures 7, 8 and 12). Based on cartographic practice, roads that do not connect to other roads or settlements should be omitted or reconnected [4]. In some cases, like in Białostocki district (Figure 7), the omission of all dead-ends would lead to the removal of many road segments, which may in turn lead to quite significant changes in road density. In such cases, reconnection of some road segments may be the solution. This issue should be further considered in future research by, for instance, specifying further topological variables and topological constraints.

In the case of the three considered districts and all considered enhanced approaches, the variables that appear on the tree are road management category (road category) and number of road connections (no. of connected roads) (see Figures 4–6). The presence of these variables at the root of the tree is logical and reasonable. The experienced cartographer would also consider them to be crucial. Moreover, by analyzing the tree from the root to its leaves, the road selection rules can be read out in a straightforward way. These rules can supplement the selection rules contained in the regulation [25]. The decision trees that are the results of DT and DT-GA in the case of 1:1,000,000 scale have the same structure at both considered detail levels. The structure of the tree is very simple, but the accuracy of the selection is high (Figure 6, Table 4). The accuracy is highest for the DT-GA enhanced approach in all tested cases. The tree contains one level only, and the decisive variable is road category. This is to be expected and is coherent with cartographic practice at the scale 1:1,000,000.

It should also be mentioned that the basic approach also yields the results that are very high in terms of the accuracy (Table 4). From this we can conclude that for very small scales, it almost does not matter which method is used. All perform very well. As Figure 6 suggests, using the road category is sufficient to select the appropriate roads. That is also the top-level rule in the regulation for the 1:1,000,000 scale (Table 1). Thus, the selection algorithms developed from DT and DT-GA for 1:1,000,000 scale will be identical. Meanwhile, the decision trees in DT and DT-GA approaches for 1:500,000 scale differ (Figures 4 and 5). The decision trees obtained for the three districts at this scale (Figures 4 and 5) are more complex in structure. For the DT-GA, the tree consists of four decisive levels, while the tree for the DT approach contains three decisive levels. We also assume that, once we

expand our analysis to include more districts, the tree may become more complex. At the same time, further careful analysis should be conducted, as an extensive decision tree may indicate an overfitting, when, instead of a rule being developed, a description of the training data is generated.

4.2. Quantitative Assessment

The roads selected in all approaches were compared against a reference map. Thus, it is possible to determine the percentage similarity based on the number of correctly selected and omitted objects. The models developed as a result of ML made it possible to improve the accuracy of selection compared to the solution applied in the basic approach. For the detail level 1:500,000, the difference in accuracy ranges from 31.51% to 49.13% in favor of the enhanced approach (Table 3). Meanwhile, for the detail level of 1:1,000,000 the difference is not so significant, though it is also in favor of the enhanced approach, ranging from 1.47% to 7.32% (Table 4).

It is worth mentioning that the accuracy obtained within the basic approach is surprisingly lower than expected (Table 4). The basic approach implements the official generalization rules contained in the regulation, thus one would anticipate the results to be of higher quality and accuracy. At the same time, for each tested district, the accuracy of the enhanced approach is higher than that of the basic approach, whether all districts are considered together or each district separately (Tables 3 and 4). The differences in accuracy are quite significant and are in favor of the enhanced approach. This thus confirms that the enhanced approach makes the selection process more efficient.

The quantitative comparison of dead-end roads also shows better results in the enhanced approach (Tables 5 and 6). For the detail of level 1:500,000 in the basic approach, there were 127 such roads, while in the case of both ML based approaches there are more than two times fewer dead-end roads, whether we consider all districts or particular one. The 1:1,000,000 scale result is the same for all approaches in case of all districts considered. In the case of the Kępiński district, one dead-end road appears in the basic approach, while in the enhanced one, no dead-ends are generated. For Rzeszowski district, in case of the basic approach, no dead-ends were generated. In each case, only one dead-end road was generated at the 1:1,000,000 scale.

Table 5. Number of dead-end roads in the result, 1:500,000 scale.

Area	Basic Approach	DT-GA	RF
All districts	127	50	56
Białostocki	51	30	31
Rzeszowski	15	7	8
Kępiński	61	13	17

Table 6. Number of dead-end roads in the result, 1:1,000,000 scale.

Area	Basic Approach	DT-GA	RF
All districts	1	1	1
Białostocki	0	0	0
Rzeszowski	0	1	1
Kępiński	1	0	0

4.3. Variable Weights

One of the important advantages of using ML models is the possibility to calculate variable weights. The weights were calculated for the road variables proposed in the enhanced approach (Table 7). The highest weight was assigned to the number of carriageways, which is in line with the rules contained in the regulation as this variable is also named as the leading one in the official documents. However, the second and third variables that were

assigned the highest weights, namely betweenness centrality and number of connected roads (segments) are new variables, proposed in the enhanced approach and not taken into account according to the regulation. Among the variables contained in the regulation, the type of road surface, road class, and road category were also assigned quite high weights, while all other variables with highest weights constitute ones that were proposed in the enhanced approach. Interestingly two variables concerning the minimum numbers of road sections leading from settlements, as well as the variable related to the connection to the settlements, were also assigned high weights. From this, we can conclude that the relations between settlements and the road network are meaningful and thus these variables should be included in the selection process. The lowest weights were obtained for the variables related to road density. These variables are also not mentioned in the regulation. The weights of all road variables are provided in Table 7. The weights of variables range from 0 to 1. In Table 7 the average weights for selected variables are presented. The weights constitute the average from the two best performing approaches, namely DT-GA and RF.

Table 7. Road variable weights.

Variable	Weight
number of carriageways	0.926
betweenness centrality	0.911
number of connected roads (segment)	0.910
minimum number of sections leading from settlements at a scale of 1:500,000	0.891
road class	0.890
number of connected roads (section)	0.883
road category	0.878
minimum number of sections leading from settlements at a scale of 1:1,000,000	0.873
connects the settlements	0.868
density of roads in the district	0.680
density of paved roads in the district	0.680
segment length	0.670
density of paved roads in a hexagon	0.668
density of roads in a hexagon	0.668

4.4. Correlation Analysis

The values of correlations between road variables were also calculated (Table 8). A strong correlation may indicate repetition of information that might warrant discarding the variable in future studies. A low correlation indicates the uniqueness of the information the variable conveys.

Table 8. Road variables correlation matrix.

Road Category	Road Class	Type of Surface	Segment Length	Density of Roads in the District	Density of Paved Roads in the District	Density of Paved Roads in Hexagon	Density of Roads in Hexagon	Betweenness Centrality	Number of Carriageways	No. of Connected Roads (Stroke)	No. of Connected Roads (Segment)	Connects the Settlements	Minimum Number of Segments Leading from Settlements at a Scale of 1:500,000	Minimum Number of Segments Leading from Settlements at a Scale of 1:1,000,000	Variables
1.00	0.54	0.44	−0.11	−0.22	−0.22	−0.31	−0.31	−0.10	−0.23	−0.27	−0.08	−0.09	−0.09	−0.07	road category
0.54	1.00	0.33	0.22	0.07	0.07	0.02	0.02	0.09	−0.25	−0.34	−0.12	0.06	−0.01	0.18	road class
0.44	0.33	1.00	−0.08	−0.12	−0.12	−0.18	−0.18	−0.02	−0.12	−0.39	−0.14	−0.02	−0.03	0.05	type of surface
−0.11	0.22	−0.08	1.00	0.86	0.86	0.87	0.87	0.25	0.00	−0.19	−0.10	0.22	0.18	0.28	segment length
−0.22	0.07	−0.12	0.86	1.00	1.00	0.81	0.81	0.22	0.01	−0.12	−0.09	0.25	0.18	0.29	density of roads in the district
−0.22	0.07	−0.12	0.86	1.00	1.00	0.81	0.81	0.22	0.01	−0.12	−0.09	0.25	0.18	0.29	density of paved roads in the district
−0.31	0.02	−0.18	0.87	0.81	0.81	1.00	1.00	0.25	0.08	−0.17	−0.11	0.26	0.22	0.35	density of paved roads in hexagon
−0.31	0.02	−0.18	0.87	0.81	0.81	1.00	1.00	0.25	0.08	−0.17	−0.11	0.26	0.22	0.35	density of roads in hexagon
−0.10	0.09	−0.02	0.25	0.22	0.22	0.25	0.25	1.00	0.05	0.00	0.12	0.12	0.11	0.14	betweenness centrality
−0.23	−0.25	−0.12	0.00	0.01	0.01	0.08	0.08	0.05	1.00	0.09	0.12	−0.05	−0.02	−0.08	number of carriageways

Table 8. *Cont.*

Road Category	Road Class	Type of Surface	Segment Length	Density of Roads in the District	Density of Paved Roads in the District	Density of Paved Roads in Hexagon	Density of Roads in Hexagon	Betweenness Centrality	Number of Carriageways	No. of Connected Roads (Stroke)	No. of Connected Roads (Segment)	Connects the Settlements	Minimum Number of Segments Leading from Settlements at a Scale of 1:500,000	Minimum Number of Segments Leading from Settlements at a Scale of 1:1,000,000	Variables
−0.08	−0.12	−0.14	−0.10	−0.09	−0.09	−0.11	−0.11	0.12	0.12	0.48	1.00	−0.19	0.00	−0.21	no. of connected roads (segment)
−0.09	0.06	−0.02	0.22	0.25	0.25	0.26	0.26	0.12	−0.05	−0.18	−0.19	1.00	0.65	0.46	connects the settlements
−0.09	−0.01	−0.03	0.18	0.18	0.18	0.22	0.22	0.11	−0.02	−0.09	0.00	0.65	1.00	0.25	minimum number of sections leading from settlements at a scale of 1:500,000
−0.07	0.18	0.05	0.28	0.29	0.29	0.35	0.35	0.14	−0.08	−0.27	−0.21	0.46	0.25	1.00	minimum number of sections leading from settlements at a scale of 1:1,000,000

Correlation

−0.39 1.00

The strongest correlations are noted between road segment length and road network density in different basic areas. The correlation between the values of these variables is logical, since the lengths of roads affect their density in the reference areas. The least correlated with other variables are technical parameters, such as type of surface and number of carriageways. These attribute variables are weakly correlated with other road characteristics and carry information about their quality and condition.

This situation indicates that in future research some of the most correlated, especially thematic attributes, could be omitted, while still maintaining high-quality results. On the other hand, spatial attributes that are least correlated with each other, as well as thematic attributes, should be retained.

5. Conclusions

In this article, we presented an innovative approach to automatic generalization of roads on small-scale maps that uses machine learning, namely decision tree (DT) models, decision tree models supported by genetic algorithms (DT-GA), and random forest (RF) models. This paper proposes variables and implements a method of the development of selection rules, data enrichment, and evaluation of the correctness of generalization for road networks. We evaluated the road generalization results in terms of qualitative and quantitative accuracy. The use of ML models made it possible to also identify the importance of the considered variables for both considered detail levels of 1:500,000 and 1:1,000,000. We also indicated variable weights and their mutual correlations. This may lead to the identification of the most crucial variables in future research, as well as allowing some of them to be excluded. However, before any such exclusion, it is necessary to enlarge the tested areas and consider more districts to analyze more representative samples.

We can conclude that the development of generalization rules supported by artificial intelligence leads to an algorithm that approximates the decision-making process previously undertaken by a cartographer. Evidence for such an outcome has already been presented in previous research on settlement generalization [13,15]. Here, we verified this assumption for an initial sample of a road selection as a representation of the network data type. The results are relevant for roads, but they might be also relevant for other utility networks such as rivers. Further analysis on other types of map thematic layers should be conducted in future research. In the enhanced approach, we obtained results that were up to nearly 50% better and closer to the maps designed by experienced cartographers than those obtained from the basic approach. The performance of the machine learning models (enhanced approach) ranges from 80.94% up to 91.23% for the 1:500,000 scale and 98.21% to 99.86% for the 1:1,000,000 scale. In contrast, in the case of the basic approach that implements very simple and limited variables, the performance is much worse, as it ranges from 42.1% to 55.25% for the 1:500,000 scale and between 91.89% and 98.39% for the 1:1,000,000 scale. It is worth noting that the difference between the basic approach and the best performing enhanced model in the case of the 1:500,000 scale, specifically for the Kępiński district, reaches nearly 50%, which constitutes a substantial improvement. For the 1:1,000,000 scale, the difference in the performance between the basic and enhanced approaches is not so significant, but still reaches over 7% in the case of the Rzeszowski district. However, in the ML results, we still see some inconsistencies and discontinuities in the road network that should be tackled in further research. For instance, although the number of dead-ends in the case of DT-GA and RF is nearly twice lower than in the case of the basic approach, for the detail level of 1:500,000, there still exist many dead-ends in the results (Table 5, Figures 7–9). While, in the case of the detail level corresponding to 1:1,000,000 scale, in both the basic and enhanced approach, the number of dead-ends is very low (Table 6, Figures 10–12). The reason for this lies in the specificities of this level of detail. At this scale only main roads are retained, which usually constitute a coherent network. Thus, in future work, the ML models used in this research should be optimized and parameterized in such a way as to ensure the overall road network consistency and connectivity. To overcome these problems, additional topological variables, for instance, further centrality

measures (closeness and degree), should be included and evaluated in the future [4,8]. For the sake of road network consistency maintenance, the topological constraints should also be defined and added in further research [28]. Moreover, the assigning of higher weights or priorities to the variables related to spatial relations between road segments and roads segments and settlements should be examined [23]. The problems related to road network consistency may be of various nature. The reason could also be the difference between source database and reference atlas map in terms of timeliness and the fact that these are two separately created datasets, based on different concepts. Thus, in future research, more optimized target datasets could be created. For instance, for the training of ML models, the output constituting the manually generalized road network GGOD data delivered by an experienced cartographer could be used.

Finally, one should note that it is not the goal of this research to reconstruct the work of a manual cartographer: after all, the manual map design process is subjective and may differ among map designers. Rather, the ultimate goal is to examine ways to help reduce the cost of map design while making the process faster and more efficient. The fact that the accuracy did not reach 100% means that further work on optimizing the road selection models is recommended. Thus, in future work, more variables and the expansion of our studies to more extensive test areas should be considered. By expanding this research, we expect to obtain decision trees that are more complex, but at the same time more informative and holistic. Valuable steps in future research would include a thorough evaluation of the achieved results with the support of experienced cartographers.

Author Contributions: Conceptualization, Izabela Karsznia, Karolina Wereszczyńska, and Robert Weibel; methodology, Izabela Karsznia and Karolina Wereszczyńska; software, Izabela Karsznia and Karolina Wereszczyńska; validation, Izabela Karsznia and Karolina Wereszczyńska; formal analysis, Karolina Wereszczyńska and Izabela Karsznia; investigation, Karolina Wereszczyńska and Izabela Karsznia; resources, Izabela Karsznia and Karolina Wereszczyńska; data curation, Izabela Karsznia and Karolina Wereszczyńska; writing—original draft preparation, Karolina Wereszczyńska and Izabela Karsznia; writing—review and editing, Izabela Karsznia, Karolina Wereszczyńska and Robert Weibel; visualization, Karolina Wereszczyńska and Izabela Karsznia; supervision, Izabela Karsznia; project administration, Izabela Karsznia; funding acquisition, Izabela Karsznia. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Centre, Poland, grant number UMO-2020/37/B/HS4/02605, "Improving Settlement and Road Network Design for Maps of Small Scales Using Artificial Intelligence and Graph Theory".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors.

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

References

1. De Serres, B.; Roy, A.G. Flow direction and branching geometry at junctions in Dendritic River Networks. *Prof. Geogr.* **1990**, *42*, 149–201. [CrossRef]
2. Yu, X. Road network simplification with knowledge-based spatial analysis. *J. Geogr. Sci.* **2001**, *11*, 54–62. [CrossRef]
3. Zhang, H.; Li, Z. Weighted ego network for forming hierarchical structure of road networks. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 255–272. [CrossRef]
4. Weiss, R.; Weibel, R. Road network selection for small-scale maps using an improved centrality-based algorithm. *J. Spat. Inf. Sci.* **2014**, *9*, 71–99. [CrossRef]
5. Benz, S.A.; Weibel, R. Road network selection for medium scales using an extended stroke-mesh combination algorithm. *Cartogr. Geogr. Inf. Sci.* **2014**, *41*, 323–339. [CrossRef]
6. Samsonov, T.E.; Krivosheina, A.M. Joint Generalization of City Points and Road Network for Smallscale Mapping. In *GIScience 2012: Seventh International Conference on Geographic Information Science*; Columbus, OH, USA, 2012. Available online: https://www.

researchgate.net/publication/264829198_Joint_generalization_of_city_points_and_road_network_for_small-scale_mapping (accessed on 5 January 2022).
7. Richardson, D.E.; Thomson, R.C. Integrating Thematic, Geometric, and Topological Information in the Generalization of Road Networks. *Cartogr. Int. J. Geogr. Inf. Geovisualization* **1996**, *33*, 75–83. [CrossRef]
8. Jiang, B.; Claramunt, C. A Structural Approach to the Model Generalization of an Urban Street Network. *GeoInformatica* **2004**, *8*, 157–171. [CrossRef]
9. Liu, X.; Zhan, B.; Ai, T. Road selection based on Voronoi diagrams and "strokes" in map generalization. *Int. J. Appl. Earth Obs. Geoinf.* **2010**, *12* (Suppl. 2), 194–202. [CrossRef]
10. Touya, G. A road network selection process based on data enrichment and structure detection. *Trans. GIS* **2010**, *14*, 595–614. [CrossRef]
11. Mackaness, W.; Beard, K. Use of Graph Theory to Support Map Generalization. *Cartogr. Geogr. Inf. Syst.* **1993**, *20*, 210–221. [CrossRef]
12. Yan, H. *Description Approaches and Automated Generalization Algorithms for Groups of Map Objects*; Springer: Singapore, 2019.
13. Karsznia, I.; Weibel, R. Improving Settlement Selection for Small-scale Maps Using Data Enrichment and Machine Learning. *Cartogr. Geogr. Inf. Sci.* **2018**, *45*, 111–127. [CrossRef]
14. Karsznia, I.; Sielicka, K. Exploring essential variables in the settlement selection for small-scale maps using machine learning. In *Abstracts of the International Cartographic Association*; Fujita, H., Ed.; International Cartographic Association: Tokio, Japan, 2019; Volume 1, p. 162. [CrossRef]
15. Karsznia, I.; Sielicka, K. When Traditional Selection Fails: How to Improve Settlement Selection for Small-Scale Maps Using Machine Learning. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 230. [CrossRef]
16. Sester, M.; Feng, Y.; Thiemann, F. Building generalization using deep learning. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-4, Proceedings of the 2018 ISPRS TC IV Mid-Term Symposium "3D Spatial Information Science—The Engine of Change", Delft, The Netherlands, 1–5 October 2018*; International Society for Photogrametry and Remote Sensing: Delft, The Netherlands, 2018.
17. Feng, Y.; Thiemann, F.; Sester, M. Learning Cartographic Building Generalization with Deep Convolutional Neural Network. *Int. J. Geo-Inf.* **2019**, *8*, 258. [CrossRef]
18. Lagrange, F.; Landras, B.; Mustiere, S. *Machine Learning Techniques for Determining Parameters of Cartographic Generalisation Algorithms*; XIXth ISPRS Congress: Amsterdam, The Netherlands, 2000; Volume XXXIII, Pt B4, pp. 718–725.
19. Balboa, J.L.G.; López, F.J.A. Generalization-oriented road line classification by means of an artificial neural network. *Geoinformatica* **2008**, *12*, 289–312. [CrossRef]
20. Zhou, Q.; Li, Z. Use of Artificial Neural Networks for Selective Omission in Updating Road Networks. *Cartogr. J.* **2014**, *51*, 38–51. [CrossRef]
21. Zheng, J.; Gao, Z.; Ma, J.; Shen, J.; Zhang, K. Deep Graph Convolutional Networks for Accurate Automatic Road Network Selection. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 768. [CrossRef]
22. Jepsen, S.T.; Jensen, C.S.; Dyhre Nielsen, T. Relational Fusion Networks: Graph Convolutional Networks for Road Networks. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 418–429. [CrossRef]
23. Gülgen, F. Road hierarchy with integration of attributes using fuzzy-AHP. *Geocarto Int.* **2014**, *29*, 688–708. [CrossRef]
24. Han, Y.; Wang, Z.; Lu, X.; Hu, B. Application of AHP to Road Selection. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 86. [CrossRef]
25. Regulation of the Ministry of Interior on 17 November 2011 on the Topographic Objects Database and General Geographic Objects Database, As Well As Standard Cartographic Products, Journal of Laws of 2011, No 279 item 1642. Available online: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20112791642 (accessed on 5 January 2022).
26. Karsznia, I.; Sielicka, K.; Weibel, R. Optimising road selection for small-scale maps using decision tree- based models. In *Abstracts of AutoCarto 23rd International Research Symposium on Cartography and GIScience Cartography and Geographic Information Society*; Redlands, CA, USA, 2020. Available online: https://tinyurl.com/58yrs79a (accessed on 5 January 2022).
27. RapidMiner 9. Operator Reference Manual 2019. Retrieved 12 May 2022. Available online: https://docs.rapidminer.com/latest/studio/operators/rapidminer-studio-operator-reference.pdf (accessed on 5 January 2022).
28. Courtial, A.; Touya, G.; Zhang, X. Constraint-Based Evaluation of Map Images Generalized by Deep Learning. *J. Geovis. Spat. Anal.* **2022**, *6*, 13. [CrossRef]