IVANA ĆAVAR, Ph.D.
E-mail: ivana.cavar@fpz.hr
ZVONKO KAVRAN, Ph.D.
E-mail: zvonko.kavran@fpz.hr
MARJANA PETROVIĆ, B.Eng.
E-mail: marjana.petrovic@fpz.hr
University of Zagreb,
Faculty of Transport and Traffic Sciences
Vukelićeva 4, HR-10000 Zagreb, Croatia

Science in Traffic and Transport
Preliminary Communication
Accepted: Feb. 25, 2011
Approved: May 17, 2011

HYBRID APPROACH FOR URBAN ROADS CLASSIFICATION BASED ON GPS TRACKS AND ROAD SUBSEGMENTS DATA

ABSTRACT

Official road classification is used for general purposes but for deep traffic analysis this classification is not sufficient. Today there are efficient ways to collect large amounts of data from multiple sources that can be used for different causes. These large amounts of data cannot be analysed with traditional methods and new state-of-the-art algorithms should be used.

The paper presents the methodology for urban road classification based on GPS (Global Positioning System) vehicle tracks and data on infrastructural characteristics of road subsegments. The process of defining road categories includes data collection and analysis, data cleansing and fusion, multiple regression, principal component analysis (PCA) as well as cross-validation and k-nearest neighbour (kNN) classification procedure. Results of such continuum can be used as base for further traffic analysis as travel time prediction, optimal route detection etc.

KEY WORDS

traffic data collection, GPS tracks, data mining, urban traffic, transportation network, road classification.

1. INTRODUCTION

In literature [1] [2] [3] [4] [5] [6] [7] [8] different road classification procedures and types can be found. Some authors [1] [2] classify roads based on their infrastructural characteristics and nominal functions. Some developed a series of updated criteria that serve for road classification or we can even find diverse classifications of roads for the same geographic area made for different purposes [4] [5] [6] [7] [9] [10] [11] [12] [13]. It is evident that traditional road classification does not meet specific needs of some advanced traffic analysis as hierarchical search of road infrastructure to determine optimal vehicle route, urban travel time prediction, etc. Therefore, there is need for new types of road classification procedures that incorporate advance algorithms and support multiple data sources.

The overall procedure starting with data collection, thought analysis and classification procedure as well as results interpretation and comparison with official classification is presented. Geographical area considered represents urban roads in the City of Zagreb.

2. DATA COLLECTION PROCESS

The process of data collection lasted for thirteen months and data collected were GPS tracks of 297 vehicles, as well as road segments data, all gathered continuously at the same urban area of City of Zagreb. The processes of data collection and cleansing, database descriptions, as well as data analysis process are presented in *Figure 1*.

In order to collect data on the traffic conditions the GPS vehicle tracks were used. The vehicles of various categories (2% of heavy vehicles and 98% of private cars) were equipped with GPS devices, and the real-time data were sent using GPRS (General Packet Radio Service), GSM (Global System for Mobile Communications) or SMS (Short Message Service) data transfer technologies to the server and then stored in the database.

The data recorded by the device and sent every travelled 100m (with the vehicle engine running) or every 5 minutes (with the vehicle engine turned off) include (*Table 1*):

- *Log time* – time of recording. Data is expressed in the number of seconds commencing from 1 January1970 (so-called Universal Time Coordinated – UTC standard);
- *Vehicle ID* – identifier of the observed vehicle / GPS device;
- *X coordinate* – x coordinate of the GPS record (WGS84 – World Geodetic System 1984);

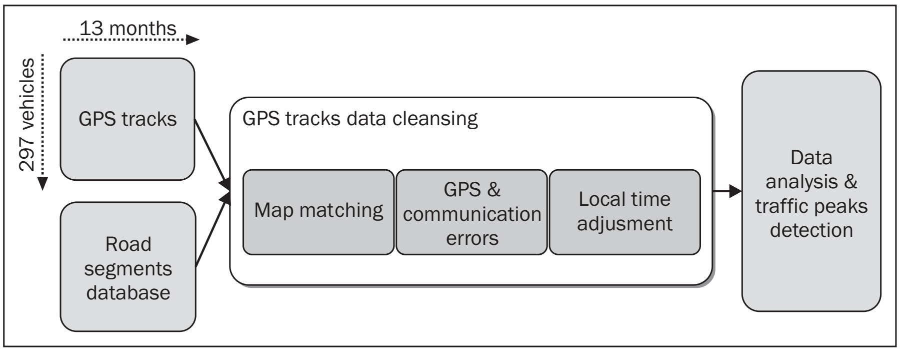

Figure 1 - Data collection process

- *Y coordinate* – y coordinate of the GPS record (WGS84);
- *Speed* – current speed in [km/h];
- *Course* – angle at which the vehicle is travelling with reference to the North;
- *GPS status* – three values that indicate the accuracy of the record. GPS status 3 indicates that the data have been collected from 4 or more satellites. GPS status 2 means that the data have been collected from 2 satellites. GPS status 1 indicates records with very questionable accuracy since the data have been collected from fewer than 2 satellites.
- *Engine status* – shows whether the vehicle engine was running or was turned off while making the recording.

Data cleansing represents the basic step in data pre-processing in order to prepare them for further analysis, and to improve the quality of the studied records. Data cleansing has been performed in three steps [13]:

Map matching

Assignment of GPS tracks to the vectorised representation of traffic network (digital map of the studied area width). A more detailed description of the digital map is given later.

Elimination of GPS errors

Encompasses the problem of GPS signal deviation from the actual vehicle location, as well as signal attenuation when the vehicle moves through a covered area or area surrounded by high buildings. Such situations very often result in a loss of signal or obtaining of results which do not correspond to the actual vehicle movement (extremely high or low speeds, etc.). In literature [15], [16] several approaches to this problem may be found. Due to a large number of vehicles and a large amount of collected data, in this paper the problem has been solved by group analysis of data. Taking into consideration that error occurs when the GPS device has low satellite visibility (for precise location at least four satellites have to be visible), all records with GPS status < 3 were eliminated. This procedure meant a loss of a large number of records, at the same time significantly increasing the quality of the remaining records.

Time variable

The collected GPS data have the time recorded in UTC format. The observed area is located in the UTC+1h time zone, but because of CEST (Central European Summer Time) a part of the year, two hours are added to UTC. The moment at which the time shifts from +1h to +2h is different every year, which required the development of an algorithm which would allocate precise local time in the observed period to the collected data.

Overall, the procedure of data cleansing removed 79% of the collected data leaving only those with satisfactory quality for further analysis.

The digital map contains all the necessary data on the roads (direction, blocked turns, marked pedestrian zones, etc.). The elements included in the database on roads are presented in *Table 2* and contain:

- *Segment ID* – identifier of the road segment;
- *Type* – numeric code of the road type according to official classification;
- *Direction* – code of the road direction;
- *Start x* – x coordinate of the beginning of the segment (WGS84);
- *Start y* – y coordinate of the beginning of the segment (WGS84);
- *End x* – x coordinate of the end of the segment (WGS84);
- *End y* – y coordinate of the end of the segment (WGS84);
- *Length* – length of the segment in metres;
- *Name* – name of the road which contains the respective segment.

The roads were divided into segments, with one segment representing a part of the road between two nodes (intersections). In the map matching procedure the vehicle tracks are projected on the digital map roads (*Figure 2*), losing in the process a part of data (either because data were outside the geographic area of interest for this paper or because the algorithm did

Table 1 - Tabular presentation of GPS records

log time	vehicle ID	x coordinate	y coordinate	speed	course	GPS status	engine status
1127839075	258	16.0311706	45.79550587	0	28	3	0
1127839076	74	16.0583437	45.71238287	69	52	3	1
1127839076	197	14.48986314	45.32554869	44	288	3	1
1127839076	80	15.94001073	45.81866081	36	336	2	1
1127839076	151	16.09944941	45.82707756	44	357	3	1
1127839077	191	16.03668246	45.80500608	0	0	3	0
1127839078	76	15.80900106	43.87035106	155	316	3	1
1127839078	2	15.68639096	43.81583756	72	139	3	1
1127839078	87	16.16276641	45.8577818	126	240	2	1
1127839078	95	18.8401363	45.33165814	37	227	3	1
1127839078	242	16.33440107	44.00933687	87	267	3	1
1127839078	21	17.11376576	45.43162725	58	325	2	1

Table 2 - Records on the road segments

segment ID	type	direction	start x	start y	end x	end y	length	name
863297106624920546	1050	2	15.911428	45.759725	15.907549	45.759386	303	Karlovačka cesta
863297106624920558	1050	2	15.975185	45.850766	15.975028	45.851058	34	Gračanska cesta
863297106624920658	1050	2	15.935334	45.767917	15.939805	45.767637	352	Karlovačka cesta
863297106624920665	1060	2	15.964125	45.752465	15.96446	45.751917	66	Sisačka cesta
863297106624920667	1060	2	15.965242	45.750619	15.96446	45.751917	156	Sisačka cesta
863297106624920681	1050	2	15.858694	45.79632	15.858315	45.796844	65	Velimira Škorpika
863297106624921008	1060	2	15.925616	45.75183	15.924161	45.750206	214	Brezovička cesta
863297106624921009	1060	2	15.925616	45.75183	15.926599	45.752731	125	Brezovička cesta
863297106624921012	1060	2	15.932676	45.739453	15.932629	45.740235	86	Dr. Luje Naletilića
863297106624921014	1060	2	15.932629	45.740235	15.932448	45.742635	267	Dr. Luje Naletilića
863297106624921065	1030	1	15.934108	45.770172	15.93363	45.770008	41	Jadranska avenija

not succeed in unambiguously allocating the record to the road segment).

Figure 2 - Digital map

In order to prepare data for further analyses it is necessary to know the basic characteristics of the collected data set and their distribution.

Prior to further analysis itself, at this point, the road data are divided to the level of subsegments. The road subsegment represents a part of segment (road section between two nodes) which has only one direction of the vehicle flow. Data prepared in this way allow separate analysis for every direction of movement and identification of traffic congestion depending not only on the congestion location but also on the flow direction itself (e.g. morning and afternoon congestions need not, and probably will not, necessarily influence both directions of the road in the same manner) [17], [18].

The number of records per subsegments ranges from one record to 63,516 records for a single subsegment and at this point there are overall 10,840,365 records. On 22 (0.17%) subsegments out of 13,236 no record has been recorded. Regarding the speed analyses on subsegments of the observed data set (*Figure 3*) we can see that arithmetic means range in the interval from 3km/h to 97.7km/h. The speeds of around 30km/h and somewhat lower (of about 20km/h), have the highest frequency, whereas the speeds higher than 55km/h occur much more rarely.

The standard deviation of speed distribution per road subsegments ranges within intervals from 0 to 36.5, with the highest frequency of about 10. The interrelation of arithmetic means and standard deviations of records in relation to the quantity of records per single subsegment is given in *Figure 4*.

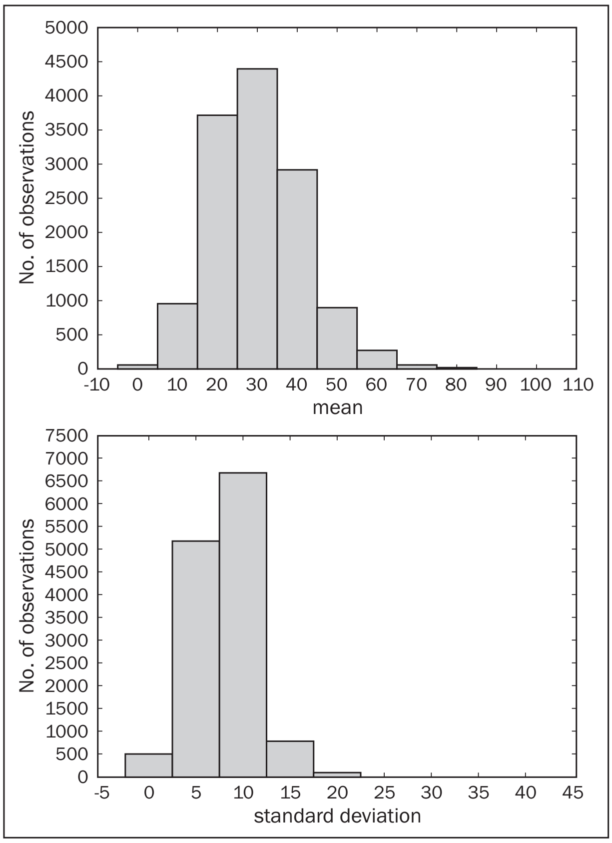

Figure 3 - Frequencies of speed records arithmetic means and standard deviations for road subsegments

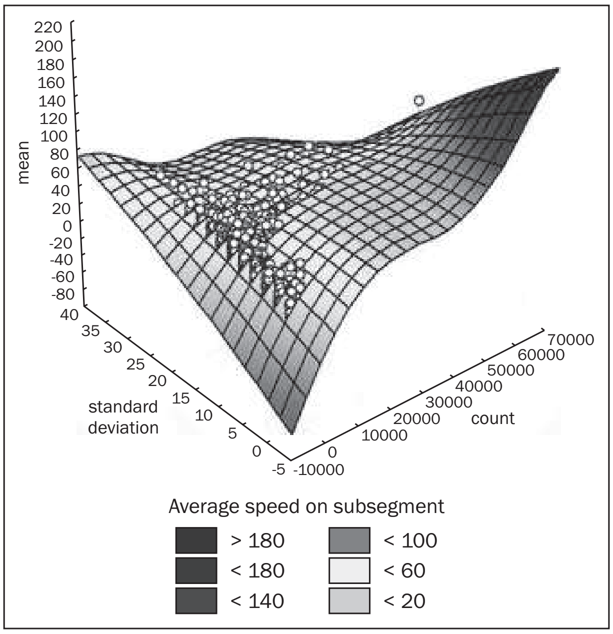

Figure 4 - 3D presentation of relation of the number of records, arithmetic means and standard deviation of GPS tracks for subsegments

The analysis of the speed records on subsegments makes it also possible to determine the time intervals in which the vehicle throughput capacity on the observed subsegments is the lowest, i.e. to identify the morning and afternoon "peak" periods. A presentation of one such analysis is given in *Figure 5* for subsegmet 4562 which corresponds to Jadranski most (the Adriatic Bridge) in the City of Zagreb for the traffic direction towards the city centre. The presentation is given for the workdays during the week, since road conditions during weekends are significantly different. The Figure clearly shows the morning congestion which is most expressed in the interval between 6 and 8 a.m. and the afternoon congestion which is a bit longer, but of lower intensity, taking from 02:30-05:30 p.m.

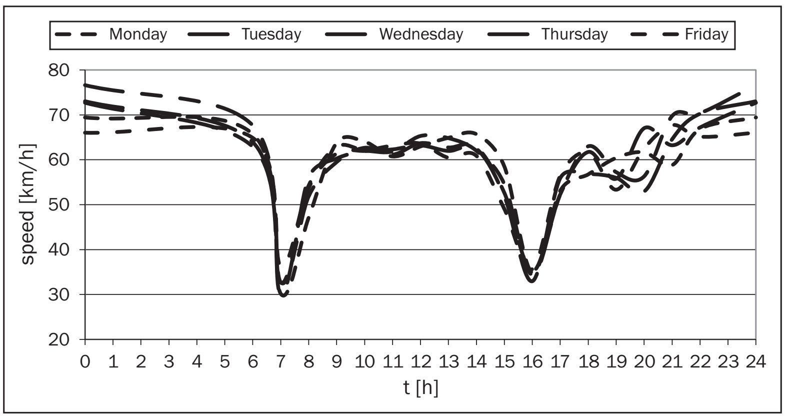

Figure 5 – Example of identifying the periods of highest traffic congestion

3. APPLICATION OF HYBRID APPROACH

Data mining is a process of automatically discovering useful information in large data repositories and represents an integral part of knowledge discovery in databases. It is a technology that blends traditional data analysis methods with advanced algorithms for processing large volumes of data. Data mining has four core tasks that include: cluster analysis, predictive modelling, association analysis and anomaly detection. Classification is a part of predictive modelling where a model for discrete target variable is built as a function of explanatory variables.

In this paper the classification methodology is used to classify road segments based on real traffic flow data collected from GPS tracks as well as road segments database described before. The overview of this process is given in *Figure 6*. The software used is StatSoft Statistica Data Miner.

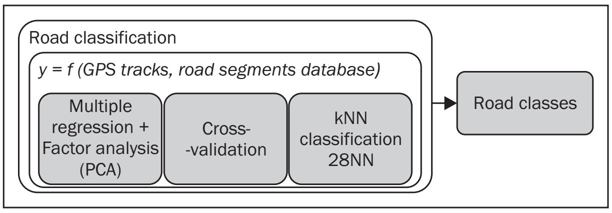

Figure 6 - Road classification process

For the purpose of road data classification a combination of multiple regression, cross-validation and *k*NN classification is used.

The *k*NN classifier represents each example as a data point in a *d*-dimensional space, where *d* is the number of attributes. Given a test sample *k*NN computes its proximity to the rest of the data points in the training set, using one of the well known proximity measures. The *k*-nearest neighbour of a given sample *z* refers to *k* points that are closest to *z* and therefore *k*NN will classify sample *z* based on the value of *k* points in its vicinity. The algorithm is highly sensitive to the choice of *k*. In fact, *k* can be regarded as one of the most important factors of the model that can strongly influence the quality of predictions [19] [20] [21]. For any given problem, a small value of *k* will lead to a large variance in predictions. Alternatively, setting *k* to a large value may lead to a large model bias. Thus, *k* should be set to a value large enough to minimize the probability of misclassification and small enough (with respect to the number of cases in the example sample) so that the *k*-nearest points are close enough to the query point. Thus, there is an optimal value for *k* that achieves the right trade-off between the bias and the variance of the model [22] [23]. In this model an estimate of *k* is provided using an algorithm known as cross-validation. Cross-validation is a well established technique that can be used to obtain estimates of model parameters that are unknown [24] [25]. The applicability of this technique to estimating *k* is described in the following section. The general idea of this method is to divide the data sample into a number of *v* folds (randomly drawn, disjointed sub-samples). For a fixed value of *k*, *k*NN model is applied to make predictions on the *v*-th segment and evaluate the error. The most common choice to measure this error is conveniently defined as the accuracy (the percentage of correctly classified cases). This process is then successively applied to all possible choices of *v*. At the end of *v* folds, the computed errors are averaged to yield a measure of the stability of the model (how well the model predicts the query points or the validity of the model for predicting unseen data). The above steps are then repeated for various *k* and the value achieving the lowest error (or the highest classification accuracy) is then selected as the optimal value for *k* (where optimal refers to cross-validation sense of optimality) [26].

The process of cross-validation is computationally very expensive; therefore, special attention should be paid to dimensionality reduction in the phase of data preparation [27] [28]. For this purpose multiple regression and factor analysis were applied. The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. In this case this was a useful way to learn about the variables and to determine what criterion is dependent and how much on other variables in the data set. On the other hand, the main applications of factor analytic techniques are to reduce the number of variables and to detect the structure in the relationships between variables. The aim of PCA is to reduce the dimensionality of a set of variables while retaining the maximum variability in terms of the variance-covariance structure. In other words, PCA tries to explain the variance-covariance structure of a data set using a new set of coordinate systems that is lesser in dimension than the number of original variables.

4. URBAN ROADS CLASSIFICATION RESULTS

The result of multiple regression and factor analysis is a set of distinctive variables among the set of GPS and road segments data:

- mean average speed on road subsegment,
- speed standard deviation on road subsegment,
- length of road subsegment,
- vehicle count of each road subsegment.

After this procedure *v* fold cross-validation is applied to determine *k* value. Cross-validation results are presented in *Figure 7*.

The optimal *k* value is determined and 28NN classification is used to classify the selected data. The distance measure used is Euclidian distance for standardised data. The data were standardised to trans-

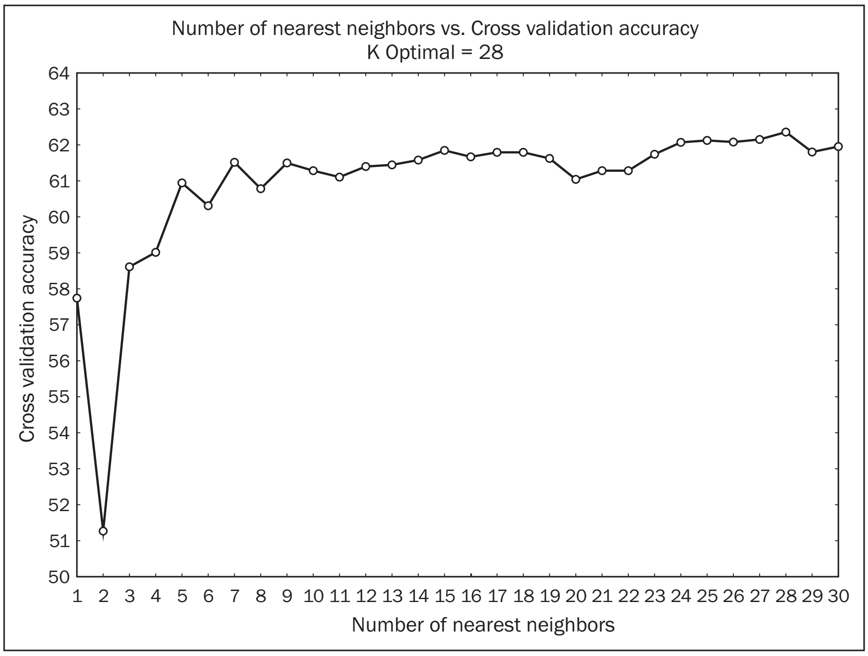

Figure 7 - Cross-validation results

form all values (regardless of their distributions and original units of measurement) to compatible units from a distribution with a mean of 0 and a standard deviation of 1. This makes the distributions of values easy to compare across variables and/or subsets and makes the results entirely independent of the ranges of values or the units of measurements.

The results of classification procedure divide roads in five classes marked as class 1, class 2, class 3, class 4 and class 5. Compared to official road classification which is explained in more detail in *Spatial planning of City of Zagreb* published by the City of Zagreb [29] (roads are divided into five road categories based on their construction characteristics and the level of authority in charge of their maintenance as highways, fast roads, state, county and local roads) these results are distinct in 47.22% cases. *Figure 8* presents the relation among official road classification (type observed), 28NN classification (type predictions) and accuracy among them. Code 101 represents class 1, code 102 class 2, code 103 class 3, code 104 class 4 and code 105 stands for class 5 of type predictions. For type observed as the 'highest' class code 101 stands for highways, 102 for fast roads, 103 for state roads, 104 for county roads and 105 for local roads.

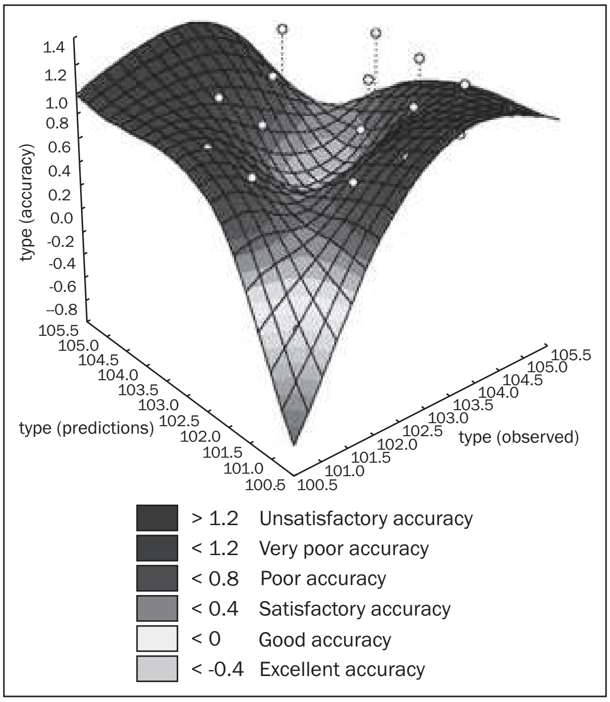

Figure 8 - Results of 28NN classification and official road classification data

The share of each class in the classification achieved by the model presented in *Figure 6* is shown in *Table 3*. The number and share of subsegments included in each class indicate that class 5 includes the largest number of subsegments (about 64%). Scrutinizing the overall length of subsegments and their share in the total road network length class 5 also has the largest share, but somewhat lower than subsegment count share. By comparison of length and subsegments shares it can be seen that class 1 has the smallest amount of subsegments, but they have a lower number of intersections and higher lengths then subsegments in other classes (length share/subsegments share for class 1 equals 3.07).

Also, 2.96% of subsegments belong to class 1, 11.76% to classes 1 and 2 and 30.36% to classes 1, 2 and 3. These classes have the highest traffic flow and 94.7% of all records were recorded there. The average

Table 3 Road classes shares

Road class	Number of subsegments	Subsegments share [%]	Length of road subsegments in class [km]	Length share [%]
Class 1	392	2.96	114,210	9.08
Class 2	1,166	8.80	175,613	13.96
Class 3	2,463	18.60	314,139	24.97
Class 4	714	5.39	67,151	5.34
Class 5	8,508	64.25	586,902	46.65

number of records for each class is given in *Table 4* where we can see that the number of records per subsegment length decreases from road class 1 to road class 5 in a non-linear manner.

Table 4 - Number of records for 1 meter of road subsegment

Road class	Number of record / / subsegment length [m]
Class 1	26.0
Class 2	15.7
Class 3	8.8
Class 4	4.1
Class 5	3.5

Some statistical measures of distinctive variables for each of the resulting classes is given in *Table 5*. The mean average of the number of records on each subsegment, vehicle mean average speed and its standard deviation as well as the average length of road subsegment decreases from road class 1 to road class 5. And while the average speed for class 1 represents double average speed of class 5 the deviation is increased. We can say that road class 1, while having the highest traffic flow and the average vehicle speed also has the largest discontinuity in vehicle movements. Respectively, class 5 has the lowest vehicle speed deviation and most continuous traffic flow.

5. CONCLUSION

The availability of new technologies and large amount of data generated in today's world joined with new state-of-the-art analysis methods can facilitate traffic analysis procedures and improve overall results within the area of traffic and transport research.

The results achieved with the application of statistical and data mining procedures as multiple regression, cross-validation and *k*NN gives road classification based on not only infrastructural characteristic of roads but on the traffic flow data as well. This allows researchers to prepare better the input data for future traffic and transport analysis and solutions by incorporating all the relevant attributes. Obstacles as large sets of multidimensional data from multiple sources are removed by dimensionality reduction based on factor analysis and multiple regression analysis facilitating computationally demanding processes as cross-validation and *k*NN classification. Such road classification allows better understanding and modelling of traffic conditions and represents good basis for more detailed traffic data analysis, as well as computer automation of this procedure. The data prepared in this way can find their application in urban travel time prediction, dynamic route guidance and/or navigation based on hierarchy algorithms, etc.

ACKNOWLEDGMENT

Authors are grateful to Mireo d.o.o. for GPS vehicle tracks and digital map provision.

Dr. sc. **IVANA ĆAVAR**
E-mail: ivana.cavar@fpz.hr
Dr. sc. **ZVONKO KAVRAN**
E-mail: zvonko.kavran@fpz.hr
MARJANA PETROVIĆ, dipl. ing.
E-mail: marjana.petrovic@fpz.hr
Sveučilište u Zagrebu, Fakultet prometnih znanosti
Vukelićeva 4, 10000 Zagreb, Hrvatska

Table 5 - Average values for all road classes

Road class	Mean average of number of records on each subsegment	Vehicle speed mean average value	Vehicle speed standard deviation	Average subsegment length [m]
Class 1	7664	62.33	15.91	292.1
Class 2	2357	49.68	14.18	150.6
Class 3	1123	42.19	12.28	127.5
Class 4	386	41.52	11.49	94.1
Class 5	240	31.00	10.01	69.0

SAŽETAK

HIBRIDNI PRISTUP KLASIFIKACIJI GRADSKIH PROMETNICA PRIMJENOM GPS TRAGOVA I PODATAKA O PODSEGMENTIMA PROMETNICA

Za potrebe dubljih prometnih analiza službene su klasifikacije prometnica često nedostatne. Danas postoje učinkoviti načini za prikupljanje velikih količina podataka iz višestrukih izvora kao i velik spektar moguće primjene ovako prikupljenih podataka. Tradicionalne se metode ne mogu primijeniti za analizu i obradu ovako prikupljenih podataka te je stoga nužno primijeniti suvremene metode za dubinsku analizu.

Rad predstavlja metodologiju za klasifikaciju gradskih prometnica primjenom podataka o GPS (engl. Global Positioning System) tragovima vozila i infrastrukturnim podatcima na razini podsegmenata prometnice. Postupak definiranja kategorija prometnica uključuje prikupljanje i analizu podataka, čišćenje i fuziju podataka, primjenu multivarijantne regresijske analize, analize glavnih komponenti (PCA), međuvalidacijske metode kao i klasifikaciju temeljem obilježja najbližih susjeda. Rezultati postignuti na ovaj način mogu biti korišteni za detaljnije prometne analize, prognoziranje vremena putovanja, utvrđivanje optimalne rute vozila i sl.

KLJUČNE RIJEČI

prikupljanje prometnih podataka, GPS tragovi vozila, dubinska analiza podataka, gradski promet, prometna mreža, klasifikacija prometnica.

LITERATURE

[1] **Chung, E.**: *Classification of traffic pattern*, 10th ITS World Congress, (CD), Madrid, Spain, 2003

[2] **Legac. I.**: *Cestovne prometnice I - Javne ceste*, Faculty of traffic and transport sciences, Zagreb, 2006

[3] Group of authors: *Transportation services - Road classification system*, City of Toronto, Canada, 2005

[4] Group of authors: *Interim PBS Road Classification Guidelines*, National Transport Commission, ARRB Transport Research Ltd, Melburne, Australia, 2004

[5] **Russ, E.R.**, **Ewy, R.**, **Mulinazzi, T.E.**, **Smith, B.L.**, **Landman, E.D.**: *Coordination of Local Road Classification With the State Highway System Classification: Impact and Clarification of Related Language in the LVR Handbook*, Kansas Department of Transportation, Topeka, Kansas, USA, 1996

[6] Group of authors: *Street Classification and Standards*, Residential Site Development Guidelines for Pennsylvania, Oversight Committee Meeting, February 10, 2005

[7] **Gill, T.**: *Development of Policy Papers for Phase Two of the Transportation Master Plan for the City of Hamilton*, Hamilton, Canada, 2005

[8] **Azam, A.**: *Road re-classification issues: Clarification on responsibility of Feeder Roads*, RIMC, LGED, Bangladesh, Pakistan, 2005

[9] **Carter, R.C.**: *Cataloging Sheet Maps*, Routledge, USA, 2003

[10] **Malone, M.**: *County road acceptance and classification*, University of Tennessee, USA, 1992

[11] **Phillips, G.**: *A road classification for use in traffic modelling*, Transport and Road Research Laboratory, USA, 1984

[12] Group of authors: *Functional classification of Washington's highway system*, Washington State Dept. of Transportation Planning Office, USA, 1992

[13] **Škorput, P.**, **Mandžuka, S.**, **Jelušić, N.**; *Real-time detection of road traffic incidents*, Promet-Traffic & Transportation, Volume 22, Issue 4, 2010

[14] **Ćavar, I.**, **Marković, H.**, **Gold, H.**: *GPS vehicles tracks data cleaning methodology*, ICTS 2006 Proceedings, Portorož, Slovenia 2006

[15] **Sarvi, M.**, **Chung, E.**, **Murakami, Y.**, **Horiguchi, R.** : *A methodology for data cleansing and trip and identification of probe vehicles*, Proceedings of the JSCE Conference of Infrastructure Planning, Vol.26, 2002

[16] **Chung, E.**, **Sarvi, M.**, **Murakami, Y.**, **Horiguchi, R.**, **Kuwahara, M.**: *Cleansing of probe car data to determine trip OD*, Proceedings of the 21st ARRB International Conference, Cairns, Australia, 2003

[17] **Bryan C. Pijanowski**, **Kimberly D. Robinson**: *Rates and patterns of land use change in the Upper Great Lakes States, USA: A framework for spatial temporal analysis*, Landscape and Urban Planning, In Press, Corrected Proof, Available online 6 May 2011.

[18] **Marković, H.** et al.; *Data-based Non-parametric Regression for Predicting Travel Times in Urban Traffic Networks*, Promet-Traffic & Transportation, Volume 22, Issue 1, 2010.

[19] **van der Walt, C.M.**: *Data measures that characterise classification problems*, University of Pretoria, South Africa, 2008

[20] **van der Walt, C.M.**, **Barnard, E.**: *Data characteristics that determine classifier performance*, Proceedings of the 16th annual symposium of the pattern recognition association of South Africa, 2006

[21] **Tan, P.N.**, **Steinbach M.**, **Kumar V.**: *Introduction to Data Mining*, Addison Wesley, USA, 2005

[22] **Pal, S.K.**, **Mitra, P.**: *Pattern Recognition Algorithms for Data Mining: Scalability, Knowledge Discovery and Soft Granular Computing*, Chapman & Hall, UK, 2004

[23] **Giudici, P.**: *Applied Data Mining*, John Wiley & Sons Ltd, Chichester, England, 2003

[24] **Wang, J.**: *Data Mining: Opportunities and Challenges*, Idea Group Publishing, USA, 2003

[25] **Wang, L.**, **Fu, X.**: *Data Mining with Computational Intelligence*, Springer, Germany, 2005

[26] **Kogan, J.**, **Teboulle, M.**, **Nicholas, C.**: *Grouping Multidimensional Data*, Springer, Germany, 2006

[27] **Larose, D. T.**: *Data mining methods and models*, John Wiley & Sons, Inc., USA, 2006

[28] **Ye, N.**: *The handbook of data mining*, Lawrence Erlbaum Associates, USA, 2003

[29] Group of authors: *Spatial planning of City of Zagreb*, City of Zagreb, Croatia, 2009