TRAVEL TIME MODELLING DATA COLLECTION

Ivana Cavar, MSc - Faculty of Transport and Traffic Sciences, University of Zagreb, Vukeliceva 4, HR-10000 Zagreb, Republic of Croatia, ivana.cavar@fpz.hr

Luka Novacko, MSc - Faculty of Transport and Traffic Sciences, University of Zagreb, Vukeliceva 4, HR-10000 Zagreb, Republic of Croatia, luka.novacko@fpz.hr

Prof. Zvonko Kavran, PhD - Faculty of Transport and Traffic Sciences, University of Zagreb, Vukeliceva 4, HR-10000 Zagreb, Republic of Croatia, zvonko.kavran@fpz.hr

ABSTRACT

The paper describes the procedures for GPS/GPRS data collection with the aim of improving the urban travel time modelling process. Data were collected from multiple sources and then fused based on common attributes. Three primary sources are GPS/GPRS vehicle data, meteorological and road infrastructure data.
GPS/GPRS data include *log_time* (time at which the record is generated expressed in UTC), *vehicle ID* (identification number of the vehicle/GPS device), *X coordinate* (X coordinate of GPS track), *Y coordinate* (Y coordinate of GPS track), *speed* (speed of the vehicle acquired from GPS in km/h), *course* (angle at which the vehicle is travelling with reference to the North), *GPS status* (3 values that indicate accuracy) and *engine status* (indicates whether the vehicle's engine was turned off or on).
Meteorological and Hydrological Service data include air and ground temperature, rain, visibility, air pressure, wind, snow etc. and GIS database infrastructure data include road link and sublink IDs as well as length, road category, direction, beginning and end coordinates, name etc.
These data were cleansed and the map-matching procedure was carried out.
All the data were collected on the same urban area during the same time period for better understanding of transportation activities. The fused data have been analysed and factors have been derived as the main carriers of input information for the travel time modelling in urban areas. The idea is to identify the main elements that affect the travel time and congestion in the urban area so they could be built into the model to achieve a satisfying level of prediction accuracy.

Keywords: data collection, GPS/GPRS, travel time modelling

1. INTRODUCTION

Travel time is a very useful and usable piece of information, interesting both for the traffic participants themselves and for the decision-makers. Besides, it represents also a comparative parameter in comparing the transport modes, and acceptable and understandable information to technical and non-technical staff in making traffic-oriented decisions. It is most frequently implemented in Advanced Traveller Information Systems (ATIS), Advanced Traffic Management System (ATMS), and for the needs of real-time vehicle guidance in urban environment, where travel time forecasting is made possible by dynamic vehicle route guidance depending on the pre-known conditions, based on the "previous experience" as well as the information on the current traffic condition.

In travel time modelling it is necessary to:
- define the main parameters that affect its variability;
- study and select methods for data collection and processing;
- consider the possibility of applying and presenting the obtained results with the aim of improving the traffic system operation.

The paper will describe all the preliminary activities necessary to prepare the data in developing one such model, starting from the very sources of data, method of collection all the way to the analysis and selection of variables that will be included in the model.

2. DATA COLLECTION

The primary concern in the process of data collection was to collect information that defines:
- the traffic condition;
- the type of road;
- the weather conditions at the moment of observation.

Consequently, the sources of data were selected, paying attention to a maximally automated process of data collection, in order to eliminate any superfluous processing while entering these into the computer model, thus obtaining a model that is fast and of higher quality.

2.1. GPS database

In order to collect data on the traffic condition the GPS vehicle tracks were used. 297 vehicles of various categories were equipped with GPS devices during thirteen months, and the real-time data were sent using wireless technologies to the server and stored in the database. This resulted in 51 835 560 vehicle tracks.
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 (UTC standard);
- *Vehicle ID* – identifier of the observed vehicle / GPS device;

- *X coordinate* – x coordinate of the GPS record (WGS84);
- *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.

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

2.1.1. GPS records database cleansing

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 [Ćavar (2006)].

- Map matching

Assigning of GPS tracks to the digital map of the studied area with vectorised traffic network. A more detailed description of the digital map is given in Chapter 2.2.

- 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. In literature [Sarvi (2002)], [Chung (2003)] several approaches to this problem may be found. Taking into consideration large number of vehicles, amount of collected data and the fact that error occurs when the GPS device has low satellite visibility, all records with GPS status < 3 were eliminated from further analysis.

- Time variable

The collected GPS data have the time recorded in UTC (Coordinated Universal Time) 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.

2.2. Digital map

The data on the roads have been taken from the digital map of the studied area. 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 segment (WGS84);
- *Start y* – Y coordinate of the beginning of segment (WGS84);
- *End x* – X coordinate of the end of segment (WGS84);
- *End y* – Y coordinate of the end of segment (WGS84);
- *Length* – length of the segment in metres;
- *Name* – name of road which contains the respective segment.

Table 2 – Records on the road segments

segment ID	type	direction	start x	start y	end x	end y	length	name
863297106624920558	1050	2	15,975185	45,850766	15,975028	45,851058	34	Gracanska cesta
863297106624920658	1050	2	15,935334	45,767917	15,939805	45,767637	352	Karlovacka cesta
863297106624920667	1060	2	15,965242	45,750619	15,96446	45,751917	156	Sisacka cesta
863297106624920681	1050	2	15,858694	45,79632	15,858315	45,796844	65	Velimira Skorpika
863297106624921008	1060	2	15,925616	45,75183	15,924161	45,750206	214	Brezovicka cesta
863297106624921012	1060	2	15,932676	45,739453	15,932629	45,740235	86	Dr. Luje Naletilica

The roads were divided into segments, with one segment representing a part of the road between two nodes (intersections). In data matching the vehicle tracks are projected on the digital map roads (Figure 1), losing in the process a part of data (either because data were out of the geographic area of interest for this paper or because the algorithm did not succeed in unambiguously allocating the record to the road segment).

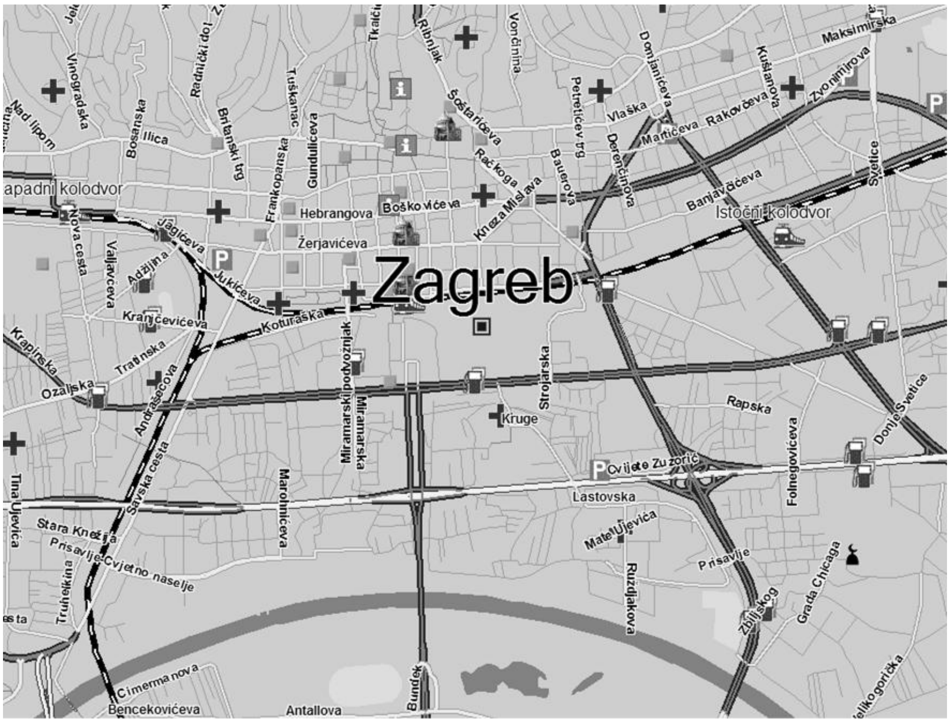

Figure 1 - Digital map

2.3. Meteorological data

Meteorological conditions play a significant role (directly or indirectly) in the traffic flow [Abdel-Aty (2005)], [Norrman, (2000)], [Eriksson (2000)], [Gustavsson (2000)] [Mathis (2000)]. The fact that the cause for about 28% of accidents on motorways and overall 19% of road accidents can be found in weather conditions [Lautenbacher (2002)] confirms this claim. Collection of meteorological data is a component of travel information and the management system defined by ITS taxonomy. In the developed ITS systems such information substantially contribute to road safety, improvement of the quality of the information provided to the users, as well as the reduction of costs (e.g. the system preventing road icing which is activated based on the meteorological forecasts on motorways results in 62% lower maintenance costs and 83% lower consumption of abrasive means and 83% less traffic accidents) [Mathis (2000)].

The data on weather conditions in the observed moment have been taken over from the authorised state institution (State Meteorological and Hydrological Service).

The weather conditions report contains the following data:

- air temperature (7-14-21h, daily mean value, daily minimal value, daily maximal value, distinction between daily maximal and daily minimal value),
- temperature of wet terrain (7-14-21h and daily mean value),
- air pressure (7-14-21h and daily mean value),
- precipitation intensity (data for 24 hours period and type of precipitation),
- s.p. (hPa) - steam pressure (7-14-21h and daily mean value),
- relative humidity (7-14-21h and daily mean value),

- cloud with phenomena (7-14-21h and daily mean value),
- wind direction (7-14-21h),
- wind intensity (7-14-21h and daily mean value),
- ground temperature at -2cm, -5cm, -10cm, -20cm, -30cm, -50cm, -100cm (7-14-21h and daily mean value),
- VV - visibility (7-14-21h),
- SS(h) - duration of sun radiance (in hours),
- E - condition of the ground (7-14-21h),
- S (cm) – overall snow cover thickness,
- Sn (cm) – new snow cover thickness.

Whereas observatory data contain data on the phenomena (rain, dew, torrential rain, haze, frost, strong wind, snow, sleet, black ice, thunder, snow on the ground, granular snow, thin hail, frozen rain, etc.) for every day during the observed period of time, and the recorded duration of the phenomenon and its intensity.

3. CHARACTERISTICS OF COLLECTED DATA AND METHODOLOGIES

In order to prepare the data for further analyses and to define the set of methods that are to be applied for their processing, it is necessary to know the basic characteristics of the collected data set and their distribution.

Prior to further analysis itself, the data in this phase are broken down 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. The 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 burden both directions of the road in the same manner).

The number of records included in the analysis, after the data preparation and cleansing process, is 10 840 365. The number of records per subsegments ranges from one record to 63,516 records for a single subsegment. 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, one can see (Figure 2 and Figure 3) 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.

Standard deviation of speed distribution per road subsegments range within intervals from 0 to 36.5, with the highest frequency of about 10. It is assumed that the traffic flow prediction on subsegments with higher standard deviation will be more difficult to be determined for a longer forecasting interval. 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.

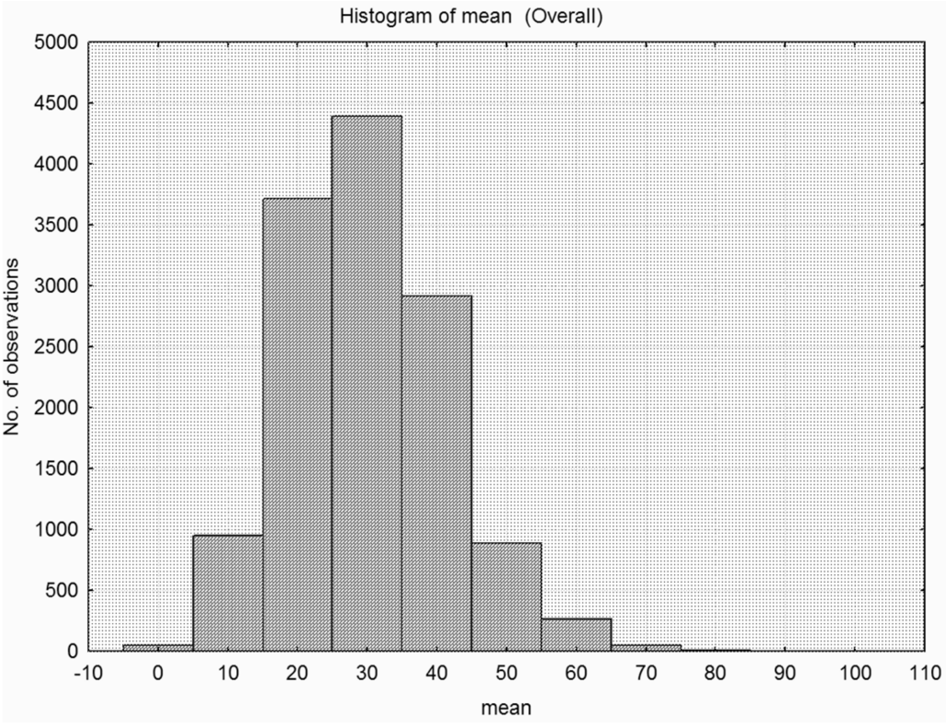

Figure 2 - Arithmetic means of speeds for road subsegments

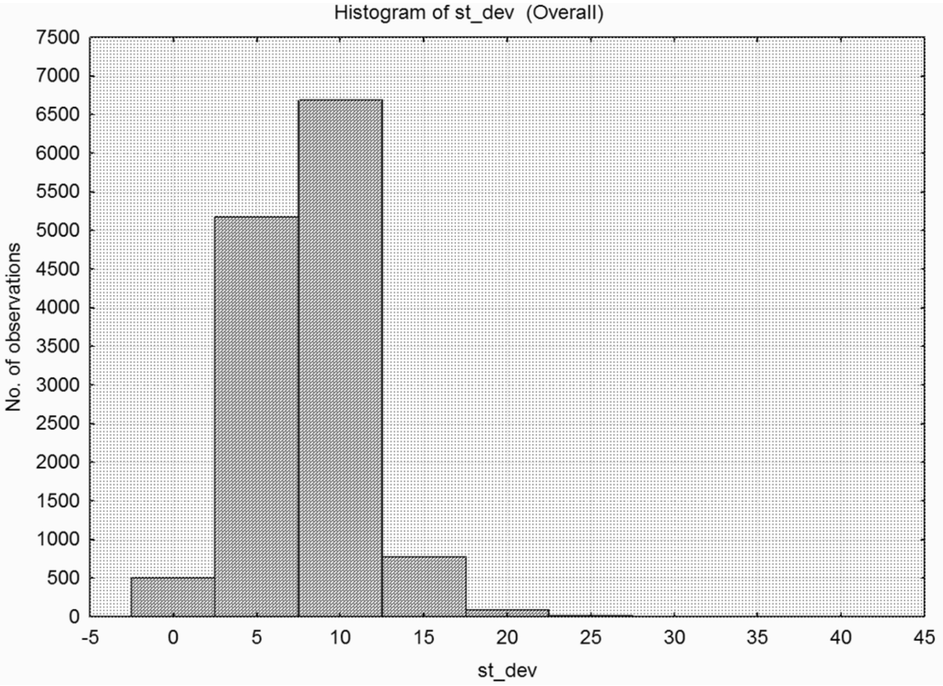

Figure 3 - Standard deviations of speeds on road subsegments

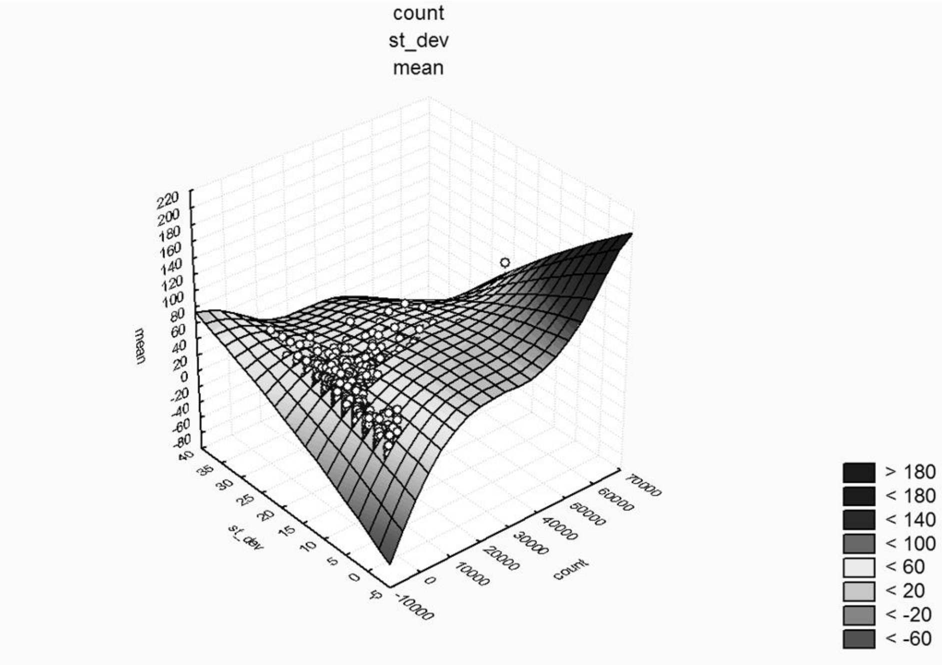

Figure 4 - 3D presentation of relation of the number of records, arithmetic means and standard deviation of GPS tracks for subsegments

Spatial distribution of vehicle tracks depending on x and y coordinates of every record can be seen in Figure 5 and Figure 6.

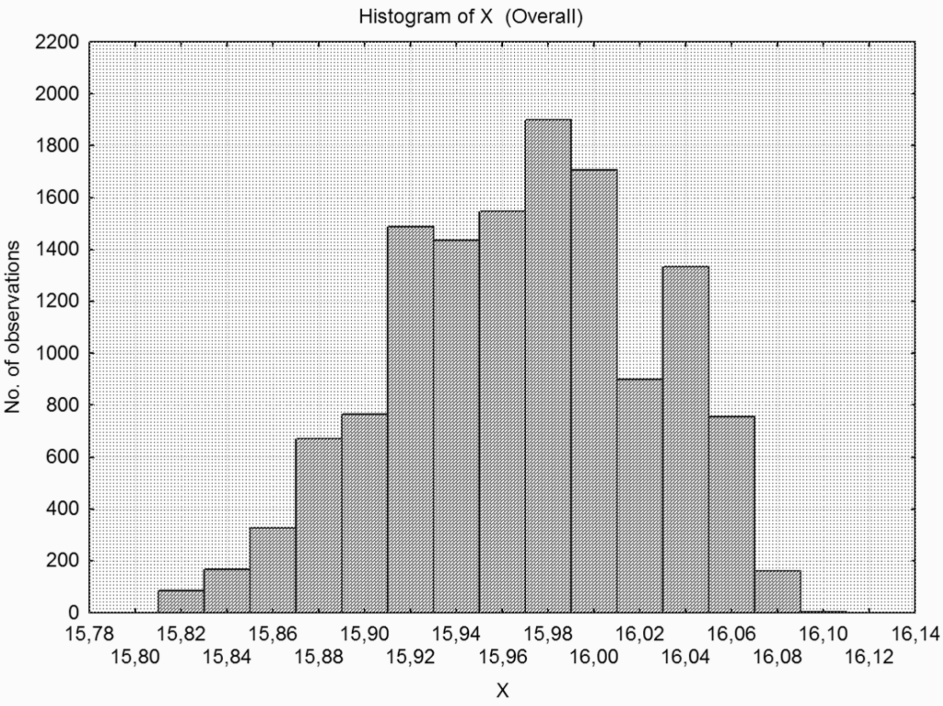

Figure 5 - Histogram of distribution of X coordinate record

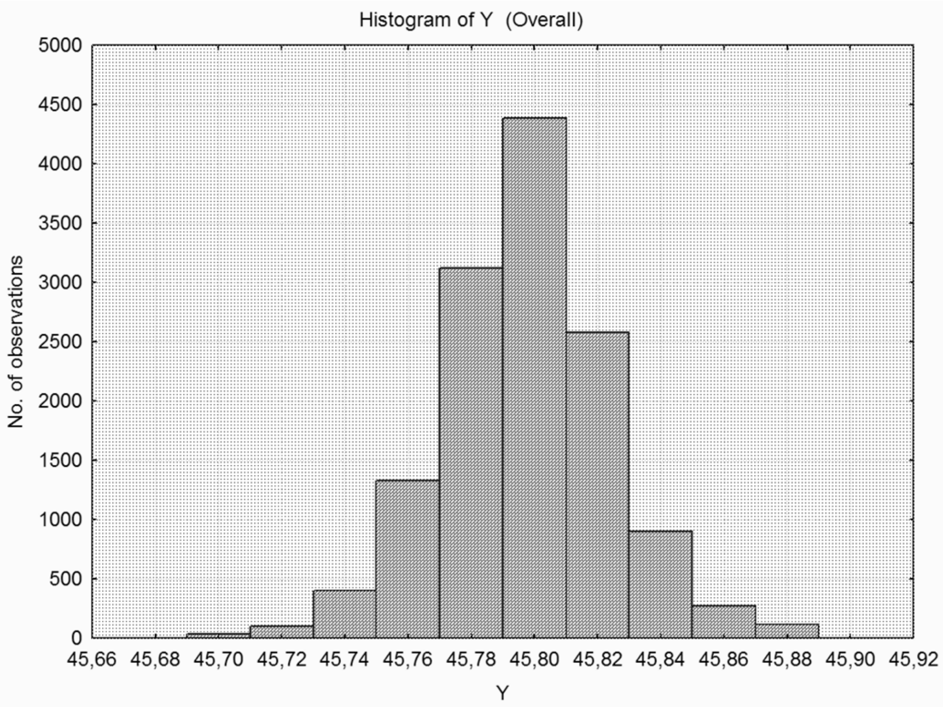

Figure 6 - Histogram of distribution of Y coordinate record

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 worst, i.e. to identify the morning and afternoon "peak" periods. A presentation of one such analysis is given in Figure 7 for subsegmet 4562 which corresponds to the Jadranski most (Adriatic Bridge) in the City of Zagreb for the traffic direction towards the city centre. The presentation is given for the working days 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 02:30-05:30p.m.

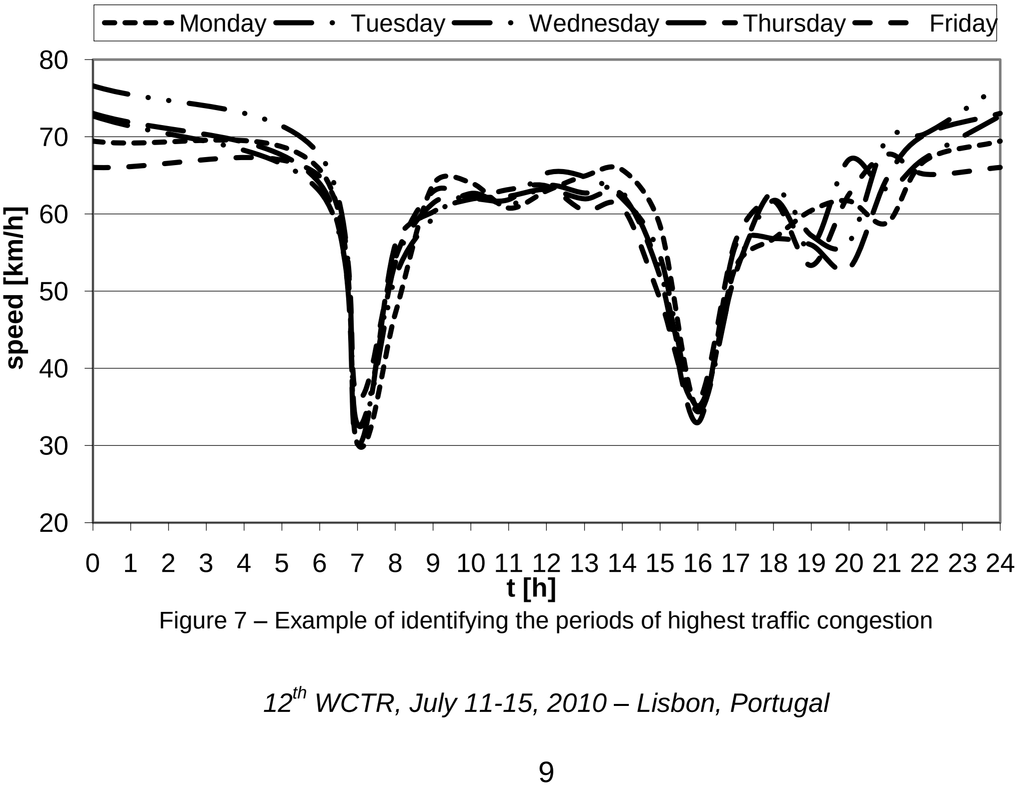

Figure 7 – Example of identifying the periods of highest traffic congestion

4. ANALYSIS OF THE JUSTIFICATION OF INCLUDING THE VARIABLES IN MODELLING

An extremely large amount of collected data represent a challenge for any type of calculation, especially for more complex analyses and/or models which should result in real-time forecasting. Therefore, a large number of records combined from three sources, with the addition of variables that are the product of data processing (arithmetic mean and standard deviation of speeds for every subsegment) represent the basis for the analysis of application efficiency in developing a model for travel time forecasting. The purpose of the analysis is to reduce the dimensions of the set of variables keeping maximal variability regarding the variance-covariance structure, i.e. to explain the variance-covariance data set structure using a new set of coordinates with smaller dimension than the number of original variables. During the analysis the interdependence of the variables is taken into consideration so that the final selection would represent the original set of variables with maximum quality.

4.1. Analysis of fused data

Analysis of fused data includes principal component analysis, multiple regression analysis and correlation analysis as shown on Figure 8.

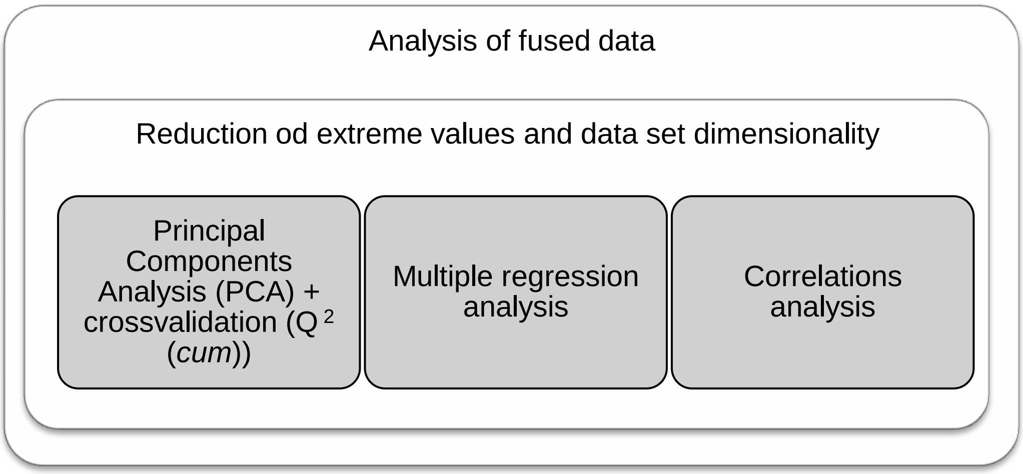
Analysis of fused data
Reduction od extreme values and data set dimensionality
Principal Components Analysis (PCA) + crossvalidation (Q 2 (*cum*))
Multiple regression analysis
Correlations analysis

Figure 8 Reduction od extreme values and data set dimensionality

The analysis of the dependence of the travel time criterion variable on the remaining 67 predictor variables including:

- Log_time, Vehicle ID, X coordinate, Y coordinate, Speed explained in Chapter 2.1;
- Segment ID, Type, Direction, Start x, Start y, End x, End y, Length explained in Chapter 2.2;
- Air temperature (7-14-21h, daily mean value, daily minimal value, daily maximal value, distinction between daily maximal and daily minimal value), temperature of wet terrain (7-14-21h and daily mean value), air pressure (7-14-21h and daily mean value), precipitation intensity (data for 24 hours period and type of precipitation), steam pressure (7-14-21h and daily mean value), relative humidity (7-14-21h and

daily mean value), cloud with phenomena (7-14-21h and daily mean value), wind direction (7-14-21h), wind intensity (7-14-21h and daily mean value), ground temperature at -5cm (7-14-21h and daily mean value), VV (7-14-21h), SS(h), E (7-14-21h), S (cm), Sn (cm) explained in Chapter 2.3.

- Count - frequency of subsegment usage,
- Mean – arithmetical mean of speeds recorded at subsegment,
- St_dev – standard deviation of speeds recorded at subsegment,
- Date –record date,
- Log_datetime – record time expressed as date and hh:mm:ss structure,

leads to the conclusion that the multiple regression equation is highly significant. R^2 of 0,58535624 (R= 0,76508577) yields the proportion of the travel time variance (fluctuation) on other variables. Also, the Durbin-Watson statistics (1.5 < d=1,867541 < 2.5) tell us that there is no systematic error or discontinuity in the data.

The variables that affect statistically significantly (α =0.05) the travel time are the time of recording, visibility, steam pressure, count, current vehicle speed, identification of the segment on which the vehicle is located, and whether it is one- or two-way road, road type, and the length of the segment between two nodes on the network. Statistically significant values and their standardized regression coefficients (Beta) and the raw regression coefficients (B) are presented in Table 3, and the intercorrelation matrix is given in Table 4.

Table 3 - Regression summary for dependent variable: travel time

	Beta	**Std.Err. - of Beta**	**B**	**Std.Err. - of B**	**p-level**
Intercept			4,491159E+13	8,440835E+12	0,000000
LOG_TIME	-32,015	7,2784	-4,08875E+00	9,295374E-01	0,000011
VV 7h	0,0415	0,0156	4,339607E-01	1,634482E-01	0,007950
VV 14h	-0,0434	0,0214	-4,246235E-01	2,094081E-01	0,042629
steam pressure (hPa) 21h	-0,0959	0,0436	-2,968157E-01	1,347613E-01	0,027664
count	0,2243	0,0866	1,151012E-02	4,445490E-03	0,009643
speed	-0,6992	0,0096	-7,693888E-01	1,060672E-02	0,000000
segmentID	-0,1163	0,0219	-5,202126E-05	9,777051E-06	0,000000
type	-0,0400	0,0157	-6,818964E-01	2,676506E-01	0,010867
direction	0,1248	0,0378	5,012279E+00	1,519459E+00	0,000977
startX	0,9515	0,2012	3,792010E+03	8,017512E+02	0,000002
endX	-0,3719	0,1138	-1,92723E+03	5,898425E+02	0,001091
length (l)	0,6233	0,1356	1,473119E-01	3,204290E-02	0,000004

Table 4 - Correlations

	VV 7h	VV 14h	s.p. (hPa) 21h	LOG TIME	count	speed	segmentID	type	direction	start X	end X	L	TT*
VV 7h	1,00	0,69	0,21	-0,43	-0,01	-0,02	0,08	-0,03	0,05	0,12	0,11	-0,04	0,02
VV 14h	0,69	1,00	0,28	-0,53	-0,01	-0,04	0,11	-0,02	0,06	0,13	0,13	-0,04	0,02
s.p. (hPa) 21h	0,20	0,28	1,00	-0,52	-0,01	-0,03	0,03	-0,01	0,07	0,17	0,17	-0,04	0,02
LOG TIME	-0,43	-0,53	0,52	1,00	0,03	0,05	0,28	0,02	0,10	0,24	0,24	0,08	-0,03
count	-0,01	-0,01	-0,01	0,04	1,00	0,36	-0,17	-0,19	0,29	-0,18	0,02	0,92	0,4
speed	-0,02	-0,04	-0,0	0,05	0,36	1,00	-0,05	-0,12	0,16	-0,25	-0,16	0,40	-0,42
segmentID	0,08	0,11	0,03	-0,28	-0,17	-0,05	1,00	0,00	0,24	-0,13	-0,15	-0,13	-0,09
type	-0,03	-0,02	-0,01	0,03	-0,19	-0,12	0,00	1,00	-0,20	-0,01	-0,03	-0,13	-0,06
direction	0,05	0,06	0,07	-0,10	0,29	0,16	0,24	-0,20	1,00	0,07	0,17	0,46	0,25
start X	0,12	0,13	0,17	-0,25	-0,18	-0,25	-0,13	-0,01	0,07	1,00	0,97	-0,39	0,03
end X	0,11	0,13	0,17	-0,24	0,02	-0,16	-0,15	-0,03	0,17	0,97	1,00	-0,2	0,1
L	-0,04	-0,04	-0,04	0,08	0,93	0,40	-0,13	-0,13	0,46	-0,39	-0,20	1,00	0,39
TT*	0,02	0,02	0,02	-0,03	0,4	-0,42	-0,09	-0,06	0,25	0,03	0,11	0,39	1,00

*TT-travel time

From principal components analysis we can see some relations among variables. Figure 9 presents loading scatter plot for principal components p1 and p2, for road Savska cesta and subsegments representing direction from the south to the north. Variables placed close to each other, as length and count, x and y, E7h, E14h and E21h, S (cm) and Sn (cm)) influence the principal components analysis model in similar ways, which also indicates they are correlated. Variables E7h, E14h, E21h, S (cm), Sn (cm) and relative humidity are positively correlated since they are positioned on the same side (quadrant) of the plot. The relative location of these variables with respect to VV21h, VV14h, VV7h, air temperature and temperature of wet terrain also indicates that they are negatively correlated to the latter.

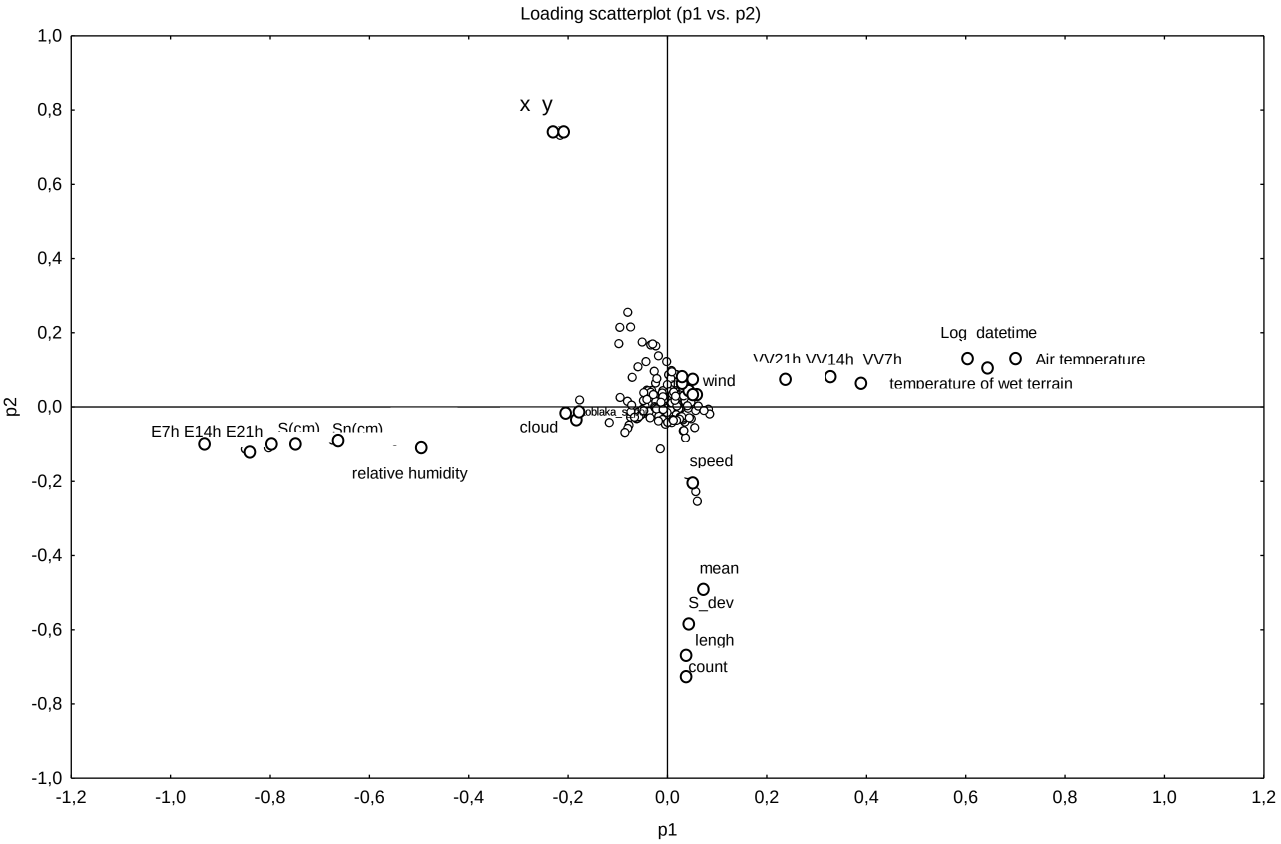

Figure 9 Principal components analysis results

4.2. Discussion on results

Variables that have high correlation, as S (cm) and Sn (cm) (r = 0,98862) and give pretty similar information on the variability of travel time, do not need to be booth included in travel time modelling process. Since roads are regularly maintained during winter conditions we can concluded that new snow cover thickness (Sn (cm)) data are sufficient. This is corroborated from the principal component analysis where we can see that these two variables influence modelling process on similar manner.

Also data that are shown to not be statistically significant in multiple regression analysis as wind intensity and wind direction, cloud with phenomena, air pressure, course etc. can be excluded from further analysis, too.

From principal component analysis one can see that variables as visibility (VV 7h, VV 14h and VV 21h) influence the model in similar ways, which also indicates that they are correlated. Multiple regression analysis shows that only two of them (VV 7h and VV 14h) are proven to be statistically significant. At the evening hours steam pressure variable seems to be more significant then measured visibility (this is due to geographical location of studied area and often occurrence of fog).

It is also interesting to note how x coordinate affects statistically significantly the travel time, which would mean that for the travel time in the observed area it is more important in which part of the city you are regarding the east-west axis than the north-south axis.

5. CONCLUSION

All three sources of the collected data justify their usability in the development of the travel time forecasting model. Besides, the data collected in this way are used simply and fast in the development of the computer model, thus enabling the development of a real-time model for travel time forecasting. The selection of the variables to be included in the development of the model does not only improve and accelerate the calculation process, but also reduces the communication costs for the real-time data transfer. Out of the available variables, only those need to be sent that have proven to be statistically significant for modelling, and the control variables that have proven their significance in the process of evaluating the quality of records simplifying the procedure of data cleansing.

By reducing the quantity of the collected and measured data by 82.09% and keeping only those that are shown to be statistically significant one can describe (R= 0,75918001) 57.64% of travel time variance. This means that 17.91% of collected records carry 98.46% of descriptive travel time variance of the entire set of data.

Also, creation of different travel time forecast models for different time intervals during the day (shown on Figure 7) is suggested. Time periods as morning congestion which is most expressed in the interval between 6 and 8 a.m. and the afternoon congestion taking from 02:30-05:30 p.m. need to be modelled more carefully than off peak periods. Due to this we can conclude that by simplifying calculation and modelling process even more, we could forecast travel time during continuous travel flow conditions adequately.

Acknowledgment

Authors of the paper would like to thank to Mireo d.o.o. and State Meteorological and Hydrological Service for data provision.

LITERATURE

Abdel-Aty,M. R.Pemmanaboina. (2005) An ATMS Accident Prediction Model using Traffic and Rain Data, 12th World Congress on ITS, San Francisco, USA.

Ćavar,I., H. Marković, H.Gold. (2006) GPS vehicles tracks data cleaning methodology, ICTS 2006 Proceedings, Portorož, Slovenia.

Chapman, A.D. (2005) Principles and Methods of Data Cleaning, Report for the Global Biodiversity Information Facility, Copenhagen.

Chung,E., M.Sarvi, Y.Murakami, R.Horiguchi, M.Kuwahara. (2003) Cleansing of Probe Car Data to Determine Trip OD, Proceedings of the 21st ARRB International Conference, Cairns, Australia.

Eriksson,D., C.Brundin. (2000) The Importance of Quality Control of Road Station Data by MESAN, Proceedings of the 10th SIRWEC Conference, Davos, Switzerland.

Gustavsson,T., J.Bogren. (2000) Speed Regulation by Use of Climate Data, Proceedings of the 10th SIRWEC Conference, Davos, Switzerland.

Lautenbacher, V.C.C. et al. (2002), Weather for Surface Transportation, National Needs Assessment Report, OFCM, Washington.

Mathis, A. (2000) Reengineering Winter Road Maintenance: Decision-making Process, Proceedings of the 10th SIRWEC Conference, Davos, Switzerland.

Norrman,J. (2000) Classification of Road Slipperiness, Proceedings of the 10th SIRWEC Conference, Davos, Switzerland.

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