Statistical modelling of traffic flow rate

Igor Grabec [1], Kurt Kalcher[2] and Franc Švegl [3]

[1] Amanova d.o.o., Technology Park, Ljubljana,
& University of Ljubljana, Slovenia
Email: igor.grabec@fs.uni-lj.si
[2] Institute for Chemistry, Karl Franzens University, Graz, Austria
Email: kurt.kalcher@uni-graz.at
[3] Slovenian National Building and Civil Engineering Institute,Ljubljana, Slovenia,
Email: franc.svegl@zag-si

ABSTRACT

The road traffic is considered as a non-autonomous dynamic phenomenon and modelled statistically by a non-parametric approach. Information for the modelling is extracted from recorded time series of traffic flow rate. The time is represented by the hour and a day-code variable specified by the calendar. An optimal predictor of the traffic flow generator is formulated in terms of conditional average estimator. The condition is comprised of hour, day-code and several data of past flow rate. As an example the traffic flow rate at a representative point on a Slovenian high-way is modelled. Seasonal variation is taken into account by using in the modelling just one month long interval of the past flow rate record. The model is utilized to forecast traffic flow rate. Applicability of the method is evaluated by the correlation coefficient r of the forecasted and original data. The mean value $\langle r \rangle \sim 0.95$ indicates rather good modelling. The performance of forecasting depends on the combination of variables representing the condition. The graph of r versus the number of condition components exhibits a maximum that determines an optimal combination of components. By the hour and day-code variables the mean traffic flow rate record over a week is determined. Its structure is described by a superposition of normal distributions whose parameters represent new information about the traffic phenomenon and activity of population.

Keywords: non-parametric statistical modelling, traffic flow forecasting, conditional average.

1. INTRODUCTION

Road traffic is a consequence of population activity to which many agents participate. In spite of this, the traffic flow does not exhibit completely random character because the population activity is synchronized to a high degree. The synchronization is stimulated externally by changing properties of the environment, as well as internally by social agreements about working days and holidays. The external stimulation can be physically described by the time and weather variables, while the internal one has to be modelled by some specific dynamic law. In agreement with these properties we consider the road traffic flow as a non-autonomous dynamic phenomenon and describe its generating equation statistically by a non-parametric model. The basic information for the creation of the model can be extracted from records of traffic flow rate and related environmental variables. For this purpose a statistical method is formulated in the next section, while in the subsequent one the applicability of the method is demonstrated by predicting the traffic flow rate at a representative point of a high-way in Slovenia. From results of this example the performance of the proposed method is quantitatively described in terms of the correlation coefficient between predicted and original values of traffic flow rate. Examination of the correlation coefficient dependence on the structure of the condition indicates how the method can be tuned to a specific case of modelling. The final goal of our approach is to provide for a quantitative forecasting of traffic flow rate that is needed for an efficient information support to participants in road traffic [2].

2. THEORETICAL BACKGROUND

The traffic flow can be quantitatively characterized by the rate $Q(t)$ of vehicles passing a certain observation point in dependence of time t [2]. Corresponding time series $\{Q(t), Q(t-1), Q(t-2),\ldots\}$ are recorded by counters at representative points of a road-network, while weather observation and forecasting services provide time series of various environmental driving variables $\{V(t), V(t-1), V(t-2),\ldots\}$. Joint time series $\{Q(t), Q(t-1),\ldots;$

V(t), V(t-1),…} comprise a data-base about complete traffic phenomenon. Since the information presented by such a data-base is usually too complex for a direct application by participants in the traffic, some pre-processing is needed. At various applications the question: "What would be the traffic load in the near future?" appears often, therefore we try to answer it by following the method of chaotic time series modelling that was previously successfully applied for prediction of energy consumption [1], [3], [5-6]. Similarly as traffic activity, the energy consumption is also a consequence of population activity, and consequently we presume that the same method would yield good results also in the case of traffic.

Driving variables *V* influencing the traffic flow usually do not depend on the flow itself and therefore we treat the flow as a non-autonomous dynamic phenomenon of chaotic character. The theory of chaotic dynamics suggests us to describe the generating process of the corresponding time series by a mapping relation [1]:

$$Q(t) = \mathrm{F}(Q(t-1), Q(t-2), ..., Q(t-\tau); V(t), V(t-1), ..., V(t-\tau)) \,. \tag{1}$$

It joins the flow rate at time *t* with the flow rate and driving variable *V* in previous times {*t*, *t*-1,*t*-2,…}. The basic problem of time series modelling is to estimate the generating function $\mathrm{F}(...)$ from a given record. For this purpose we first estimate the probability density function (PDF) of measured variables and use it to express the generating function by an optimal statistical predictor [1]. With this aim we describe the traffic by the state vector **S** = (*Q(t),Q(t-1),…,Q(t-τ);V(t), V(t-1),…,V(t-τ)*) that joins the flow rate *Q* at various moments with the corresponding driving variable *V*. We consider **S** as a random vector variable whose properties can be generally characterized by some probability density function f(**S**). For this purpose we extract from given records *N* representative samples {**S***n* ; *n=1,…, N*}. Based on them we express f(**S**) by the Parzen's kernel estimator [4]

$$\mathrm{f}(\mathbf{S}) = \frac{1}{N} \sum_{n=1}^{N} w(\mathbf{S} - \mathbf{S}_n, \sigma) \tag{2}$$

The kernel $w(\mathbf{S} - \mathbf{S}n, \sigma)$ denotes an approximation of the delta function, for example the Gaussian or the Lorenzian one, while σ represents the distance between samples **S***n*.

Assume next that some component *P* of the state vector depends on the remaining components **R** of **S**. We proceed to the estimation of the corresponding functional relation *P*(**R**) statistically by following the concept of minimal estimation error [1]. The resulting optimal statistical estimator is the conditional average:

$$\hat{P}(\mathbf{R}) = \mathrm{E}[P \mid \mathbf{R}] \tag{3}$$

E[|] denotes the conditional mean value and **R** the condition. After expressing the conditional average by the probability density function Eq. (2) , we express the estimator $\hat{P}(\mathbf{R})$ in terms of samples **S***n* = (P*n* , **R***n*) as:

$$\hat{P}(\mathbf{R}) = \frac{\sum_{n=1}^{N} P_n \, w(\mathbf{R} - \mathbf{R}_n, \sigma)}{\sum_{i=1}^{N} w(\mathbf{R} - \mathbf{R}_i, \sigma)} \tag{4}$$

This estimator has already been applied for modelling of various chaotic time series and corresponds to a normalized radial-basis function neural network [1], [3]. In order to use it in modelling of traffic flow rate, we consider again Eq. (1) and interpret variables on its right side as the condition:

$$\mathbf{R} = (Q(t-1), Q(t-2), ..., Q(t-\tau); V(t), V(t-1), ..., V(t-\tau)) \,, \tag{5}$$

while the variable on the left side as the value of the flow rate *P=Q(t)* which we want to predict. Equations (3) & (4) are then readily applicable for this purpose. However, at a selected time *t* we have to provide the variables comprising the condition that is represented by the truncated sate vector **R**. If we want to forecast a sequence of flow rate in the more distant future, we can repeat the complete procedure for the next time step with the forecasted value included into the condition and with new values of other variables comprising the condition.

Before the application of the proposed method we must specify the dimension of the state vector **S** by the value of parameter τ . Since its is generally not known in advance how many past values have to be utilized in modelling, we can proceed to a proper value of τ by observing the performance of the forecasting at increasing values of τ . For this purpose it is of advantage to describe the performance quantitatively by the correlation coefficient *r* between predicted and measured time series of the traffic flow rate.

2. EXAMPLE OF MODELLING AND FORECASTING

For our demonstration we utilize records of traffic flow rate collected by automatic counters on Slovenian roads in one hour time intervals over the year 2007 and published by the Slovenian Roads Agency on a CD: ISSN-1580-3864. As a representative example we arbitrary selected the record from the counter 822 on a high-way from Ljubljana to Postojna. The record presents data about physical time of measurements and flow rate of various categories of vehicles. The time was transformed to a periodic variable that is uniformly increasing from 0 to 24 over each day. More demanding is a proper transformation of time to a proper day-variable. Analysis of electrical power and natural gas consumption has revealed [3], [5], [6] that phenomena depending on population activity essentially depend on the character of the day which we also consider here as the driving variable of the traffic. We describe the character of the day quantitatively by a code defined by the following rule: Monday – 1, day after holiday or weekend – 2, normal working day – 3, Friday – 5, day before holiday or weekend – 6, Saturday – 7, Sunday – 9, holiday – 10. The resulting distribution of the day-code variable over the year 2007 is shown in Fig. 1. Among various categories of vehicles we consider the category 'personal cars' since it is the most numerous. A record of its flow rate is shown in Fig. 2.

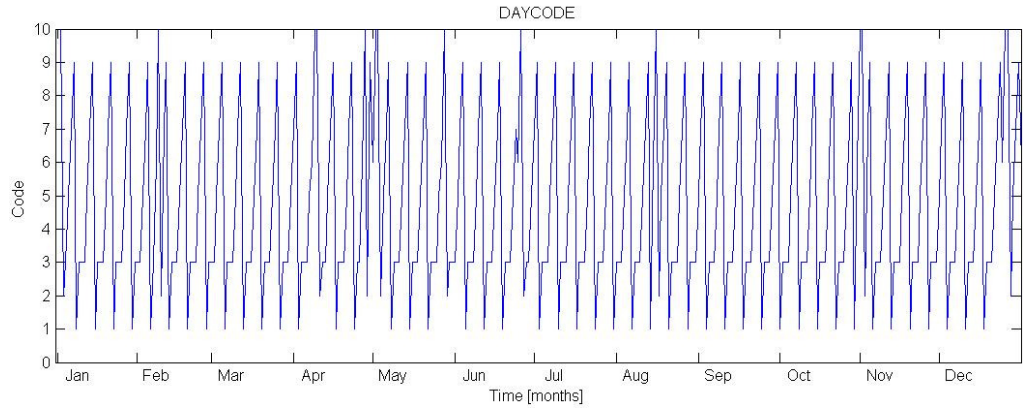

Figure 1. Distribution of the day-code in the year 2007.

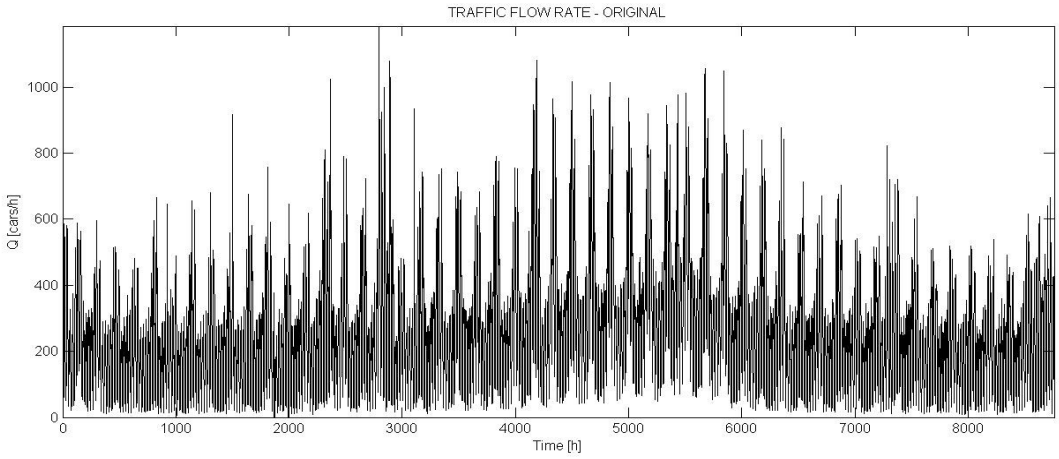

Figure 2. Record of the traffic flow rate *Q(t)* in the year 2007.

The record reveals rather regular seasonal variation of flow rate over the year. The traffic activity culminates in the summer due to contribution of tourists travelling on the high-way from Ljubljana to the Adriatic seaside. In

the modelling, stemming from a one-year record, the influence of seasonal variation can be accounted for by forming the model based upon shorter intervals. In our treatment we use a record spanning one month.

Beside seasonal variation, the record exhibits rather regular variation of traffic flow in normal working days and rather irregular variations in days around holidays. The most outstanding irregularities are observed around May 1st, November 1st and the end of a year, when many people try to join weekends with holidays by going on leave. To demonstrate this remarkable property we consider here two characteristic examples that correspond to a normal week and the week around holidays. The first one is the week No. 16 spanning from April 16 to 22, while the second one is the week No. 18 from April 30 to May 6 which includes two holidays. Graphs of all characteristic variables are shown in Fig. 3 for both cases. When forecasting the flow rate shown in these graphs, the condition in the model was comprised from the day-code and the hour variable alone.

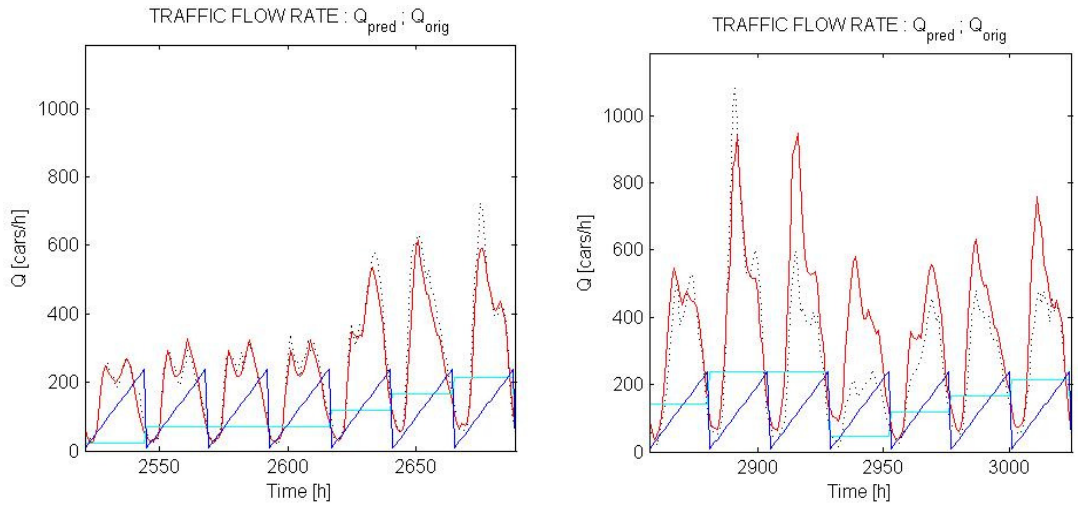

Figure 3. Graphs of characteristic variables in a normal week No.16 (left) and in a week No.18 (right) that includes May 1st and 2nd. The step-like and saw-like curves represent the day-code and the hour. The dotted curve shows the original, while the solid one shows the corresponding predicted record of traffic flow rate.

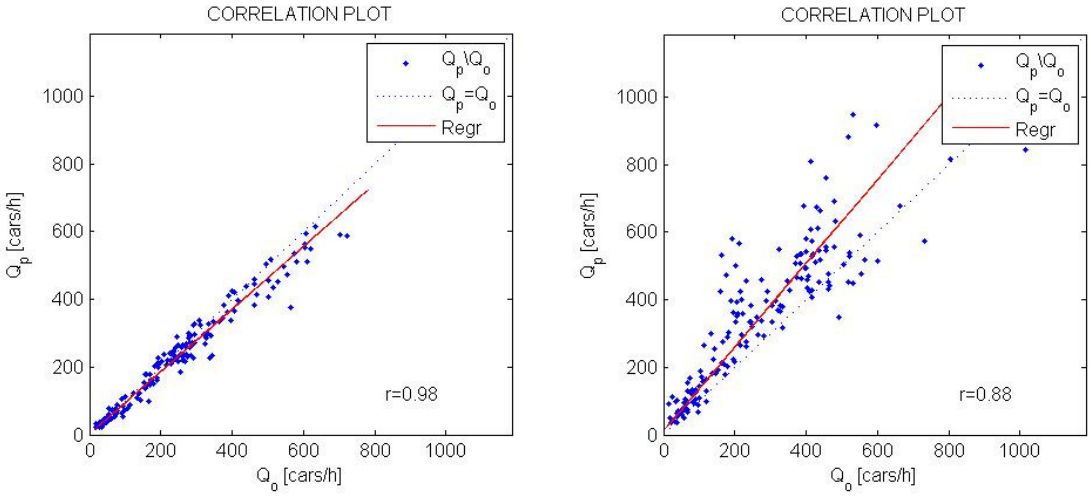

Figure 4. Correlation plots of predicted *Qp* versus original *Qo* flow rate corresponding to the normal week (left) and the week with holidays (right).

4. DESCRIPTION OF THE MODELLING PERFORMANCE

The agreement between predicted and original records of traffic flow rate given in Fig. 3 is quantitatively demonstrated by the correlation plots in Fig. 4. The horizontal and vertical axes represent original *Qo* and predicted *Qp* flow rate respectively. To each hour of a week there corresponds a point (*Qo,Qp*) in the graph. From the distribution of points the correlation coefficient *r* and the linear regression line (solid) were determined.

The dotted line represents an ideal agreement between original and predicted data. The value of correlation coefficient r and the agreement between regression and dotted line indicates the quality of forecasting. The values r = 0.98 and r=0.88 correspond to the normal week and the week with included holidays, respectively. The main reason for a worse prediction in the week with embedded holidays is an absence of a similar holiday in the data-base used in modelling of the flow generator. Research of energy consumption has shown that the prediction during holidays could be improved by using records from various years [3], [5-6].

The presented examples correspond to rather extreme properties of traffic flow. In order to demonstrate the mean performance of forecasting, we use the plot of correlation coefficient over the year that is shown on the left side of Fig. 5. The corresponding mean value $\langle r \rangle$=0.94 indicates that the modelling and forecasting is on average quite successful. However, the prediction is rather good in "normal" days while it is worse during holidays. We conjecture that the influence of weather conditions is much more expressed during holidays, since then the population activity is mainly determined by travelling, sporting and relaxation engagements of the population. Therefore, we suppose that the traffic flow rate in a short time interval before the prediction exhibits the influence of existing weather, and expect that the performance of the forecasting could be improved by including the corresponding data into the condition.

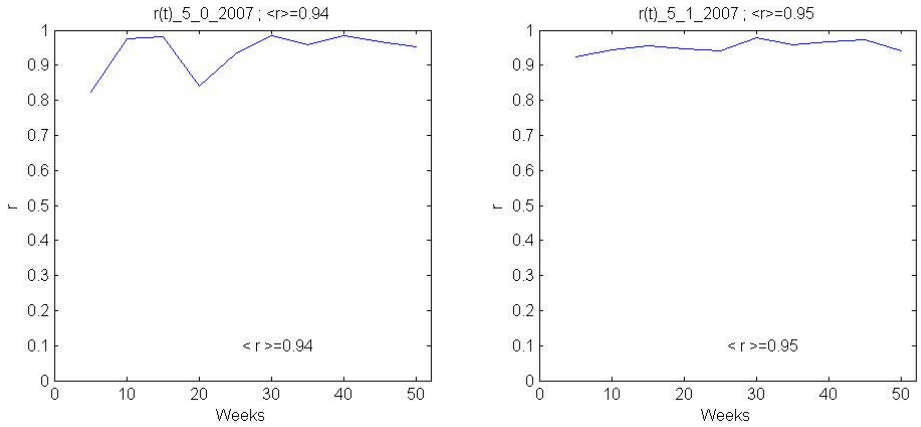

Figure 5. The correlation coefficient between predicted and original flow rate over the year 2007. The left record is obtained by using just the day-code and hour variables as the condition, while the right one is obtained by considering also the value of flow rate before the prediction.

In order to prove our conjecture we have changed the composition of the condition **R** by including into it also several past components of traffic flow Q. By analysing the forecasting performance of models with various numbers of included past flow rate values, we have found that the best model is obtained by including just the final value of traffic flow rate before the prediction. For this case the dependence of the correlation coefficient on time in the year is shown by the right graph in Fig. 5. The corresponding mean value $\langle r \rangle$=0.95 is not significantly higher in this case, but it is of advantage that the correlation coefficient fluctuates less over the year than in the previous case. It is a bit surprising, that inclusion of more past data of flow rate again diminishes the performance. This effect is shown in Fig. 6.

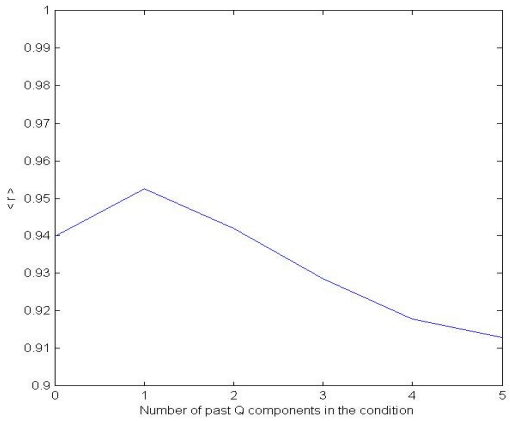

Figure 6. Dependence of the correlation coefficient on the number of past flow rate values in the condition.

4. MODELLING OF TRAFFIC FLOW RATE BY NORMAL DISTRIBUTIONS

Introduction of the hour and the day-code variables provides a proper tool for the analysis of traffic phenomenon and related population activity. Using both variables as the condition, we can simply extract from an arbitrary selected time interval of a given data base the mean flow rate for each hour and type of the day. By composing corresponding data we obtain a mean record of flow rate over a week. Fig. 7 shows two examples of such records corresponding to a "normal week" without holidays. The record on the right side represents the traffic at the previously mentioned point on the high-way in Slovenia, while the right one represents the traffic in the vicinity of Helsinki in Finland.

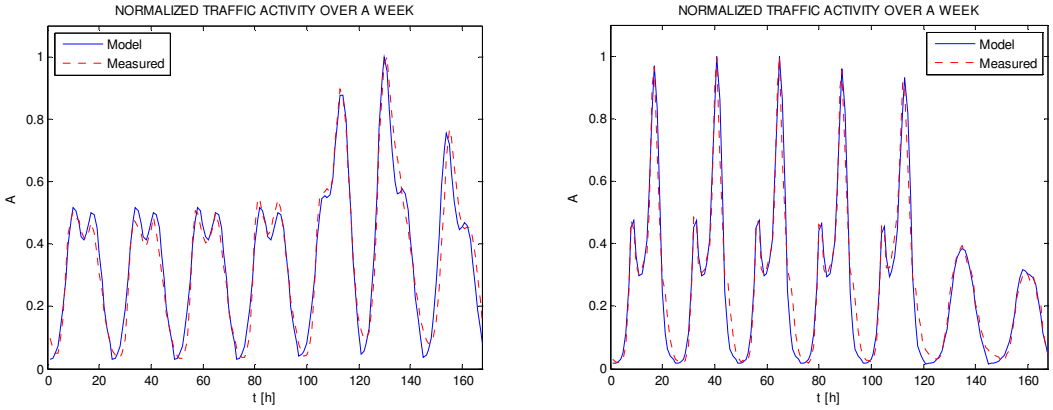

Figure 7. Distribution of a normalized mean traffic flow rate in normal weeks of the year 2007 determined from data from Slovenia (left) and Finland (right). The normalized traffic activity is defined as: $A=<Q(t)>/<Q(t)>_{max}$. Dotted line - original data, and the solid line - data determined by a normal superposition model.

In spite of an expressive difference between both records, their structure exhibits a remarkable common property. In both cases the complex distribution of the flow rate over each particular day indicates that it is comprised of several more elementary distributions resembling a normal one. Based upon this observation we have tried to describe analytically the distribution in a particular day by a superposition of normal (Gaussian) distributions:

$$Q_d(t) = \sum_{k=1}^{K} A_{dk}\,\mathrm{G}(t - T_{dk}; S_{dk}) + C \qquad (6)$$

Parameters A_{dk}, T_{dk}, S_{dk} represent the amplitude, the mean value and the standard deviation of the k-th Gaussian distribution component in the day indicated by index d, and K denotes the number of normal components comprising the superposition, while C is an additive constant. All parameters can be adapted mathematically by minimizing the discrepancy between the superposition of Eq. (6) and the actual distribution. However, few attempts are needed to determine them approximately based upon visual inspection of a given record. According to this, we have found the following parameters of the superposition for both characteristic examples.

4.1 Parameters of normal superposition for data from Slovenia

Parameters are: T_1=9.9h, T_2=17.6h, S_1=2.7h, S_2=3.1h, C=0.03 are constant while amplitudes are variable, as shown in Table 1.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
A_1	1.0	1.0	1.0	1.0	1.0	2.0	1.5
A_2	1.0	1.0	1.0	1.0	1.8	1.1	0.9

Table 1. Amplitudes of normal superposition for data from Slovenia.

4.2 Parameters of normal superposition for data from Finland

Parameters S_3=4.0h and C=0.015, are constant while the other ones are variable, as shown in Table 2.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
A_0	0.32	0.32	0.32	0.32	0.32	0.35	0.22
A_1	0.35	0.35	0.35	0.34	0.33	0.05	0.08
A_2	0.72	0.75	0.75	0.68	0.71	0.02	0.12
T_1	8.40	8.40	8.40	8.40	8.40	12.0	13.5
T_2	17.2	17.2	17.2	17.2	17.2	17.5	19.0
T_3	14.0	14.0	14.0	14.0	14.0	15.3	15.0
S_1	1.25	1.25	1.25	1.25	1.25	1.65	1.65
S_2	1.50	1.50	1.50	1.50	1.70	2.30	2.30

Table 2. Parameters of normal superposition for data from Finland.

It is characteristic the data from Slovenia could be quite well described by using in each day just two normal components with equal amplitudes and widths, and different mean values in the morning or in the afternoon. The parameters of these components are approximately constant for working days from Monday to Thursday, while they differ for Friday, Saturday and Sunday. In Fig. 7 the original data are shown by the dotted line, while the data determined by the superposition are shown by a solid one. Similarly, the records from Finland could be described, however with different parameters. In this case we have found that an additional normal component with mean value at noon significantly contributes to the accuracy of the model. In this case the parameters of three components in the days from Monday to Friday are approximately constant, while they differ in Saturday and Sunday. In both cases the constant C does not represent a significant contribution.

Rather good agreement of original and modelled data indicates that the normal superposition quite successfully represents some basic properties of observed traffic flows. We conjecture that each normal component of the superposition represents a synchronized activity of some group of population and expect that the corresponding characteristic parameters could be also determined based upon a proper sociological research.

4. CONCLUSIONS

Our goal was to develop a simple statistical method for modelling and forecasting of traffic flow rate. We have found that the conditional average estimator is applicable for this purpose since it does not require any analytical modelling of traffic flow. Consequently, the complete structure of the corresponding model stems from recorded data. The developed method is therefore rather generally applicable. For instance, we have obtained similar prediction performance as on data from Slovenia also on data obtained from Finland. The main problem at the application of the proposed method is to describe the character of the day by a proper code. In spite of diminished prediction performance caused by less predictable population activity during holidays, the mean value of correlation coefficient $\langle r \rangle \sim 0.95$ indicates that the developed statistical method is quite applicable for forecasting. Beside rather good forecasting, the introduced driving variables that represent the hour and day-code also render possible a simple description of traffic flow by a superposition of normal distributions. Parameters of components in the superposition can be applied for a reduced representation of traffic flow dynamics, while a proper interpretation of this property still awaits more profound sociological examination of population activity.

It is clear that the population activity is not synchronized just by the hour and character of the day, but also by the environmental conditions determined by variations of weather. Until now we have tried to account this effect by including the past flow rate into the condition. But the proposed method permits rather simple inclusion of weather variables directly into the condition of the model as well. Even more, weather conditions predicted by various weather observation services could provide for still more reliable forecasting of traffic flow in the near future. Therefore, this possibility is now further investigated.

5. REFERENCES

[1] Grabec, I. and Sachse, W. 1997. *Synergetics of Measurements, Prediction, and Control,* Springer, Berlin.

[2] Kerner, B. S. 2004. *The Physics of Traffic*, Springer, Berlin.

[3] Lunar, B. and Grabec, I. 2002. *Forecasting Electrical Power Consumption by Normalised Radial Basis Function Neural Network*, Neural Network World, 12(3):241-254

[4] Parzen, E. 1962. *On Estimation of Probability Density Function and Mode*, Ann. Math. Stat., 35:1065-1076.

[5] Potočnik, P., Thaler, M., Govekar, E., Grabec, I. and Poredoš, A. 2007. *Forecasting risks of natural gas consumption in Slovenia*, Energy policy, 35(8):4271-4282.

[6] Thaler, M., Grabec, I. and Poredoš, A. 2005. *Prediction of Energy Consumption and Risk of Excess Demand in a Distribution System*, Physica A: Stat. Mech. & Applications, 355(1):46-53.

ACKNOWLEGEMENT

This research was financially supported by the project Roadidea from the EU 7FP. The authors would like to thank the Slovenian Agency for Roads, and especially Mrs. Tatjana Bubnič for preparation of flow rate data. Data from Finland were provided by Ari Sirkiä from VTT – Technical Research Centre of Finland.