Metodološki zvezki, Vol. 7, No. 1, 2010, 25-37

Optimal Compression of Traffic Flow Data

Igor Grabec[1]

Abstract

Experimental characterization of complex physical laws by probability density function of measured data is treated. For this purpose we introduce a statistical Gaussian mixture model comprised of representative data and probabilities related to them. To develop an algorithm for adaptation of representative data to measured ones we introduce the model cost function by the sum of discrepancy and redundancy. All statistics are expressed by the information entropy. An iterative method is proposed for searching the minimum of the cost function that yields an optimal model. Since representative data are generally less numerous than measured ones, the proposed method is applicable for compression of overwhelming experimental data measured by automatic data-acquisition systems. Such a compression is demonstrated on the characterization of traffic flow rate on the Slovenian roads network. The flow rate during a particular day at an observation point is described by a vector comprised of 24 components. The set of 365 vectors measured in one year is optimally compressed to just 4 representative vectors and related probabilities. These vectors represent the flow rate in normal working days and weekends or holidays, while the related probabilities correspond to the relative frequencies of these days. However, the number of representative data depends on the accuracy of PDF estimation.

Keywords: data compression, Gaussian mixture model, cost function, traffic flow.

PACS: 06.20.DK - Measurement and error theory, 02.50.+s - Probability theory, stochastic processes, and statistics, 89.70.+c - Information science.

1 Introduction

Road traffic is a very complex stochastic phenomenon, whose basic properties have to be specified by measurements (Helbing, 1997; Kerner, 2004). Due to its complexity measured data are overwhelming and there generally appears a question how to process or archive them efficiently. Merely by intuition, one should suggest to use for this purpose a set of representative data that carry *a proper amount of information*. However, the problem is to specify this amount

[1] Amanova Ltd, Technology Park 18, 1000 Ljubljana, Slovenia; igor.grabec@amanova.si

so that it could be estimated automatically in an information processing system. We expect that the corresponding specification would provide a basis for a development of a computer code by which representative prototype data could be created from measured ones. The basic request is that prototype data should represent approximately the same information about the phenomenon as the measured ones but should be less numerous.

It is known that evolution of human intelligence is related to formation of notions that are utilized in thought processes. The notions are primarily formed by a self-organized interaction of neurons in the brain based on excitation by very complex signals generated by a network of sensors and in fact represent our perceptions in a kind of compressed thought representation of the world. We therefore expect that methods developed in the field of artificial intelligence in relation to self-organized memory formation could lead us to an efficient compression of overwhelming traffic data to acceptable representative ones (Grabec et al., 1997; Grabec, 1990; Kohonen, 1989). For this purpose we utilize methods of experimental statistical modelling of physical laws developed recently (Grabec et al., 1997; Grabec, 2001, 2005). In the next chapter we explain the fundamentals of this modelling, which leads us to the formulation of the generalized Gaussian mixture model and its cost function. The adaptation of the model to measured data that follows from the minimization of the cost function finally leads us to a proper specification of an optimal set of representative data. For this purpose a new method is developed here and applied to the formulation of an optimal model of traffic flow in terms of representative data. Its performance is demonstrated on traffic data recorded on the roads network in Slovenia.

2 Fundamentals

Let us consider a phenomenon characterized by N measurements of a variable x using an instrument with span $S_x = (-L, L)$. Properties of the instrument are specified by calibration on a unit u. The PDF of the instrument's output scattering during calibration is described by the scattering function $g(x,m)$. When the scattering is caused by mutually independent disturbances in the instrument, the scattering function can be assumed to be Gaussian (Grabec et al., 1997; Lesurf, 2002):

$$g(x,\bar{x}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x-\bar{x})^2}{2\sigma^2}\right] \tag{2.1}$$

We apply this function in our further treatment. The mean value $\bar{x}$ and standard deviation σ can be estimated statistically by calibration performed on the measurement instrument using the unit u. In addition to this we introduce a uniform reference PDF: $\rho(x) = 1/(2L)$ indicating that all outcomes of the experiment are hypothetically equally probable before executing an experiment.

Let x_i denote the most probable instrument output in the *i*-th experiment. Using $g(x,x_i)$ we describe the properties of the explored phenomenon during the *i*-th experiment. Similarly, the properties in a series of *N* repeated experiments, which yield the basic data set $(x_n ; n=1,...,N)$, are described by the experimental PDF over the kernel estimator:

$$f_N(x)=\frac{1}{N}\sum_{n=1}^{N} g(x,x_n) \tag{2.2}$$

Properties of this estimator have been widely examined elsewhere (Parzen, 1962; Fukunaga,1972; Duda et al., 1973; Révésh, 1991; Grabec et al. ,1997; Grabec, 2001, 2005) and could be summarized as follows. With an increasing number of experimental data *N*, the experimentally estimated PDF approaches the hypothetical one and consequently the kernel estimator Eq.(2.2) resembles a consistent estimator. Although it is biased, the bias can be diminished if the accuracy of experimental observation is improved by decreasing the standard deviation σ. Hence, the kernel estimator appears to represent a solid basis for an acceptable experimental estimation of PDF. However, a deficiency is that the formulation of the kernel estimator does not include a specification of a proper number of experimental data, and consequently, we try to improve it by the definition of a generalized Gaussian mixture model (Fukunaga, 1972; Duda et al., 1973; Révésh, 1991; http://en.wikipedia.org/wiki/Mixture_Model; Grabec et al., 1997).

3 Gaussian mixture model and representative data

During the experimental acquisition of data it is usually observed that initial data points are on average well separated. In this case the scattering functions in the sum of Eq.(2.2) are not overlapping and it seems reasonable to keep them in the PDF estimator. However, with a very large number of data *N*, a new datum on average falls into close vicinity of some previously acquired one which results in the overlapping of scattering functions, and consequently, it seems redundant to include a new term into the estimator. A particular data sample contributes to the empirically estimated PDF at the sample point an amount that is proportional to its probability weight $p=1/N$. This further means that overlapping results in an ever finer adaptation of the experimentally estimated PDF to the hypothetical one. However, a similar effect can be achieved by just adapting the probability weight *p* of already existing data instead of increasing the number of terms in the estimator. This reasoning leads us to a slight generalization of the PDF estimator by changing Eq.(2.2) to the Gaussian mixture model equation:

$$f_M(x) = \sum_{m=1}^{M} p_m \, g(x,q_m) \tag{3.1}$$

Here the couple p_m, q_m denotes the probability p_m and the centre q_m of the *m*-th *representative datum*. *M* such couples represent the parameters of the PDF model. To simplify our expressions we further use for the model basis functions the notation $g(x,q_m) = g_m(x)$.

With the introduction of the model, there immediately appears a question how could we judiciously determine a proper set of its parameters. At the first glance the problem could be classified as the adaptation of a parametric model to some hypothetical probability density function *f*. But this is in fact much more involved since we do not know in advance neither the number *M* of representative data nor the values $(p_m, q_m ; m = 1,...,M)$. In addition, we also do not know the hypothetical distribution *f* but just its empirical estimator given by Eq.(2.2). So a proper strategy for the solution of this problem has first to be found. For this purpose we introduce the model cost function.

In order to formulate the cost function we temporarily assume that the hypothetical probability density function $f(x)$, which characterizes the phenomenon, is given, while later on we shall describe it experimentally by the kernel estimator given by Eq.(2.2). Our goal is to adapt the model PDF $f_M(x)$ to the hypothetical $f(x)$. For this purpose we first describe their discrepancy (*ie* the estimation error) by the Kullback-Leibner divergence (Grabec et al. 1997; Cover et al.,1991; Kolmogorov, 1956; MacKay, 2003):

$$D = \int_{S_x} [f_M(x) - f(x)] \log \frac{f_M(x)}{f(x)} \, dx \tag{3.2}$$

Our goal is to diminish the discrepancy by adaptation of representative data. However, we want to employ in the model just a proper number of these data. With this aim we proceed with the definition of model information and redundancy similarly as in previous articles (Grabec, 2001, 2005, 2009a,b). For this purpose we describe the indeterminacy by the negative value of the relative entropy:

$$H_M = -\int_{S_x} f_M(x) \log \left[\frac{f_M(x)}{\rho(x)} \right] \, dx \tag{3.3}$$

We next define the entropy of model basis functions using means over model probabilities p_m:

$$H_b = -\int_{S_x} \sum_{m=1}^{M} p_m \, g_m(x) \log \frac{g_m(x)}{\rho(x)} \, dx \tag{3.4}$$

Using the difference of these two statistics we define the *model information*:

$$I_M = H_M - H_b \tag{3.5}$$

Its maximal possible value is $I_{M,\max} = \log M$, and therefore the model redundancy is defined as:

$$R_M = \log M - I_M \tag{3.6}$$

The model cost function is then introduced by the sum of discrepancy and redundancy:

$$C_M = D_M + R_M \tag{3.7}$$

In the case when the model closely fits the hypothetical PDF the model divergence D_M can be approximately expressed as $D_M \approx -I_M + const.$. This expression then yields: $C_M \approx -I_M + R_M + const.$ The statistic $C = -I_M + R_M$ can be interpreted as the *information cost function* (Grabec, 2001, 2005), and therefore the model cost function is in this case approximately determined by the information cost function: $C_M \approx C + const.$. When proceeding to the strategy for model adaptation we may therefore take into account the properties of the information cost function explained in details elsewhere (Grabec, 2001, 2005).

4 Adaptation of mixture model parameters to experimental data

If we want to adapt the model Eq.(3.1) to an arbitrary hypothetical PDF $f(x)$ we must specify the number M and values of parameters $(p_m, q_m; m = 1,...,M)$ that yield the minimum of the model cost function. We cannot perform this by a standard variation method since M is an integer number (Grabec et al, 1997). Even if we know this number, equations derived by variation of cost function are non-linear and cannot be solved easily. Consequently we do not consider the general problem anymore, but rather, turn to the problem of model adaptation to an increasing number of experimental data values. Various approximate methods of growing and pruning have been proposed for this purpose in the field of neural networks (Haykin, 1999; Leonardis et al., 1998). The growing methods are mainly utilized when the model is adapted to an increasing number of experimental samples, while pruning is used when a large number of experimental samples is compressed to a smaller number of representative data. Since we do not want to store overwhelming traffic data we next consider just the case with the increasing

number of experimental samples. In the literature various criteria have already been proposed for this purpose, but at present there is still no generally accepted method (Haykin, 1999; Leonardis et al., 1998). Consequently, we here propose a new method that is based upon minimization of the model cost function stemming from experimental information.

At the adaptation of the model to an increasing number of data we first substitute the hypothetical probability density function $f(x)$ by its experimental kernel estimator $f_N(x)$ specified by Eq.(2.2) and then try to adapt $f_M(x)$ to $f_N(x)$. In order to proceed to a proper strategy for the solution of the adaptation problem let us consider the case when $f_M(x)$ is well adapted to $f_N(x)$: $f_M(x) \approx f_N(x)$ and examine what happens when the number of experimental data values is increased from N to $N+1$. In this case the PDF estimator $f_N(x)$ is changed to $f_{N+1}(x)$. The corresponding difference is:

$$\Delta f_N(x) = f_{N+1}(x) - f_N(x) = \frac{1}{N+1} g_{N+1}(x) - \frac{1}{N+1} f_N(x) \qquad (4.1)$$

The first term represents a localized contribution of the $N+1^{st}$ data sample, while the second one represents the decrease of the complete estimator in order to preserve the normalization of PDF. Since $f_M(x) \approx f_N(x)$ we can most simply adapt the model to the last term of Eq.(4.1) by multiplying all probabilities $p_m; m = 1,...,M$ by the factor $1 - 1/(N+1) = N/(N+1)$. Less direct is adaptation to the first term on the right side of Eq.(4.10). For this purpose let us consider two characteristic possibilities:

I. The model already contains a prototype q_c that takes place in a close vicinity of the $N+1^{st}$ data sample x_{N+1} so that $|q_c - x_{N+1}| << \sigma$. In this case we can achieve a proper adaptation by joining the new sample x_{N+1} and the closest prototype q_c. At this step the representative probability is increased: $p_c \to p_c + 1/(N+1)$ and its centre is moved to the centre of gravity of joined terms: $q_c \to \frac{q_c p_c + x_{N+1}/(N+1)}{p_c + 1/(N+1)}$. As the weight for calculation of centre of gravity we use the probability appertaining to the prototype and the sample value.

II. All prototypes take place far away from the $N+1^{st}$ data sample: $|q_m - x_{N+1}| >> \sigma$ for all *m*. In this case we can achieve a proper adaptation by creating a new representative by the sample value: $q_{M+1} = x_{N+1}$ and probability $p_{M+1} = 1/(N+1)$.

We have found by numerical investigations (Grabec, 2009 a,b) that in both cases the model cost function is minimally changed. If the model is well adapted to N data, then its cost function is approximately minimal. Since the addition of the new sample changes it for approximately minimal possible amount, the cost function of the model adapted to $N+1$ samples by the corresponding step I or II is again approximately minimal one and we can expect that the estimation $f_M(x) \approx f_{N+1}(x)$ is still valid.

It is instructive to examine what in fact happens if we follow the above described procedure of adaptation. In the first case the number of prototypes M and $\log M$ are preserved, the model information is minimally changed and consequently also the redundancy as well as the cost function are minimally changed. In the second case, the number of representative data M and $\log M$ are increased, but the corresponding change of model cost function is properly counterbalanced by the increased experimental information so that the cost function minimum is again approximately preserved.

The adaptation steps described in I and II can be simply transformed into a computer code. A transition from the first to the second possibility occurs when the distance between a new sample and the closest representative surpasses the width σ of the model basis function. The resulting procedure starts with equalizing the first data sample with the first prototype: $x_1 = q_1$ and yields an approximately optimal model. If there is given an overwhelming set of experimental data samples, the corresponding Gaussian scattering functions in the experimental estimator of PDF Eq.(2.2) exhibit expressive overlapping which corresponds to a high information cost. In such a case we can expect that the corresponding optimal model would be comprised of essentially less representative data whose basis functions would be just slightly overlapping, while the resulting PDF estimator would approximately correspond to the hypothetical PDF. The proposed method thus represents an efficient compression of data without essential loss of information, and is therefore proposed for the treatment of traffic data before their preservation in memory units.

At the application of the proposed method there appears a problem when the width σ of the scattering function is not known. In this case the value of σ should be specified by the user and with respect to the wanted accuracy of PDF modelling. Quite generally one can suggest to normalize the variable x with respect to its variance. This yields a relative variable: $x_r = x / \sqrt{\mathrm{var}(x)}$. Specification of σ_r that corresponds to this variable then determines how detailed is the modelling of the PDF. Such a treatment also renders possible comparison of various models. A problem also appears when the phenomenon is characterized by a vector variable $\mathbf{x} = (x_1, x_2, ..., x_d)$. In this case one can suggest to normalize i-th component separately with respect to its variance: $x_{r,i} = x_i / \sqrt{\mathrm{var}(x_i)}$. If the components represent similar properties of the phenomenon, then some common

average σ can be applied as a relative measure for the description of the level separating the steps I and II of the method. In our subsequent example of traffic flow modelling we apply this method.

In relation to the formulation of the mixture model given in Eq.(3.1) one could expect that there is possible still further generalization by adapting also the width of basis functions $\sigma \to \sigma_m$. Although such a generalization could yield still more compressed presentation of the PDF (Kohonen,1989), the interpretation of the model basis functions in terms of scattering functions is lost, and consequently a representative datum cannot be interpreted as a typical result of a measurement. For an application in the framework of traffic data analysis this is not convenient, and consequently, we do not proceed here with this next step of generalization.

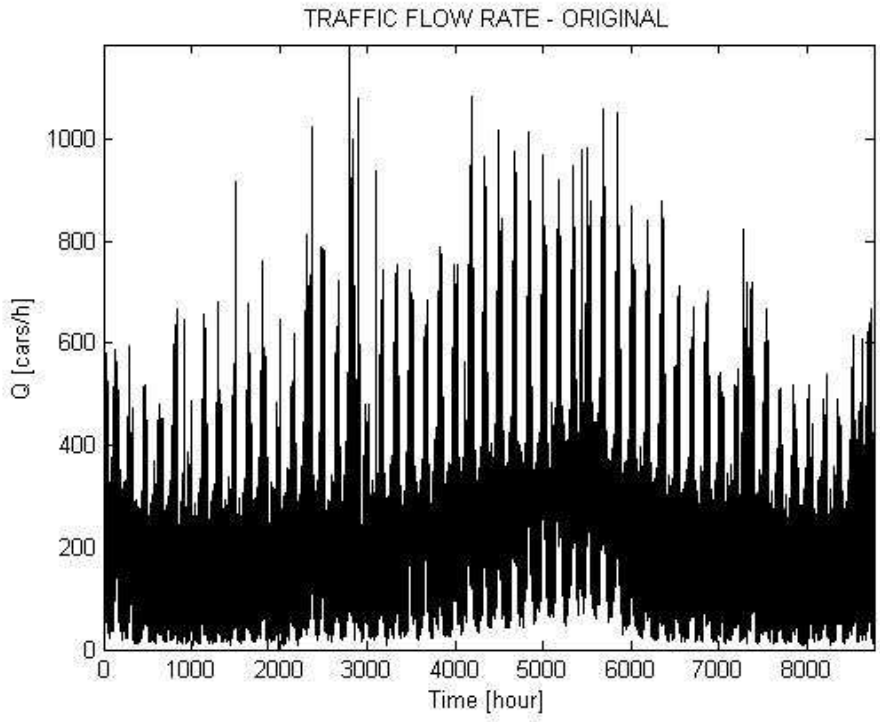

Figure 1: An example of traffic flow record from a Slovenian road in the year 2007.

5 Example of the optimal traffic flow model

In order to demonstrate the applicability of the proposed optimal modelling we consider observation of traffic flow rate at a certain point of road. As an example we utilize here data obtained in the year 2007 from a counter 822 at some representative point on a high-way in Slovenia and published by the Slovenian Roads Agency on a CD as the file: PROMET_ASP_QLD6_2007_STM822.xls. For the sake of simplicity we consider just a single category of cars and describe

the measured traffic flow rate by the variable Q. This variable depends on time of observation t so that a measurement yields a record of corresponding time series: $\{Q(t);t=0,\Delta t,2\Delta t,...,T\}$. Here Δt denotes the length of time interval between successive measurements. In our treatment we normally use one hour intervals: $\Delta t = 1\text{h}$ which we also adopt here. An example of the record spanning over the year 2007 is shown in Figure 1. Without loss of generality for the description of traffic phenomenon, we further decompose the recorded time series into a set of traffic flow day-vectors that are comprised of data recorded over a day d : $\mathbf{Q}(d) = (Q_1(d),...,Q_{24}(d))$. The day-vectors extracted from the record in Figure 1 are shown by green lines in Figure 2 together with the mean day-vector denoted by (-*-). As the index of traffic day-vector component we use the hour of its recording.

In order to proceed to the modelling we must specify some criterion for the separation of the steps I and II of the adaptation process. For this purpose we calculated variance of each day-vector component using statistical average over all days of the year. Using the corresponding variances we introduced for each component the relative distance between a given day-vector and a representative one as: $d_i = (Q_i - q_i)/\sqrt{\text{var}(Q_i)}$. Using it we introduced a measure of discrepancy between the day-vector and the representative one by the mean square relative distance over a day: $\delta = \sum_{i=1}^{24} d_i^2 / 24$. As a separation level between the steps I and II we then used the value $\delta_s = 2$. This value has been selected based upon numerical testing of modelling performance in which we have been looking for a characteristic level that provides for a formation of prototypes corresponding to characteristic days of the week.

Using the day-vectors and the level of separation $\delta_s = 2$ we formed the representative vectors as described by steps I and II. The corresponding representative vectors are shown by black lines in Figure 2 as well as separately in Figure 3. In order to demonstrate applicability of representative vectors we calculated from the model PDF the mean day vector. Its record is shown in Figure 3 by the line (-o-) and coincides well with the record of the mean vector calculated directly from the complete set of day-vectors (-*-). This coincidence confirms a correct representation of traffic variable PDF by the model.

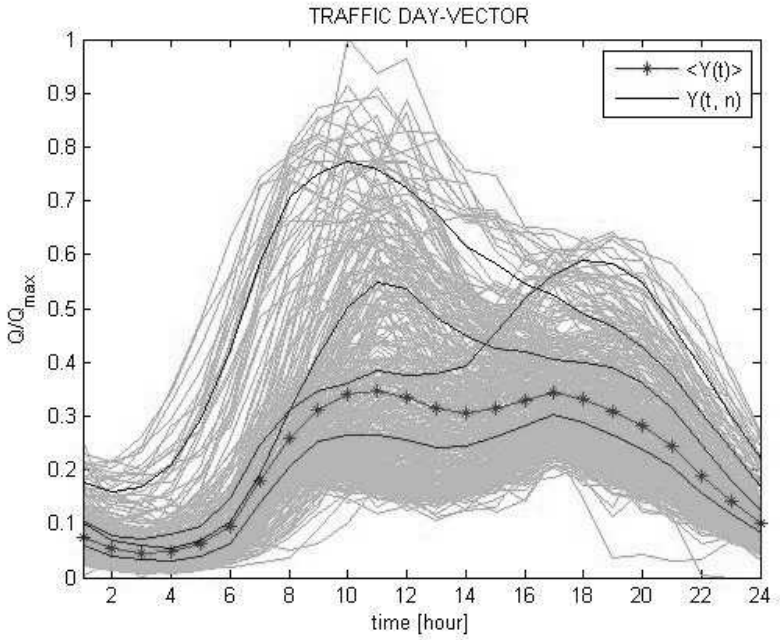

Figure 2: Traffic day-vectors formed from traffic flow record presented in Figure 1 (green) representative vectors (black) and the mean vector (-*-).

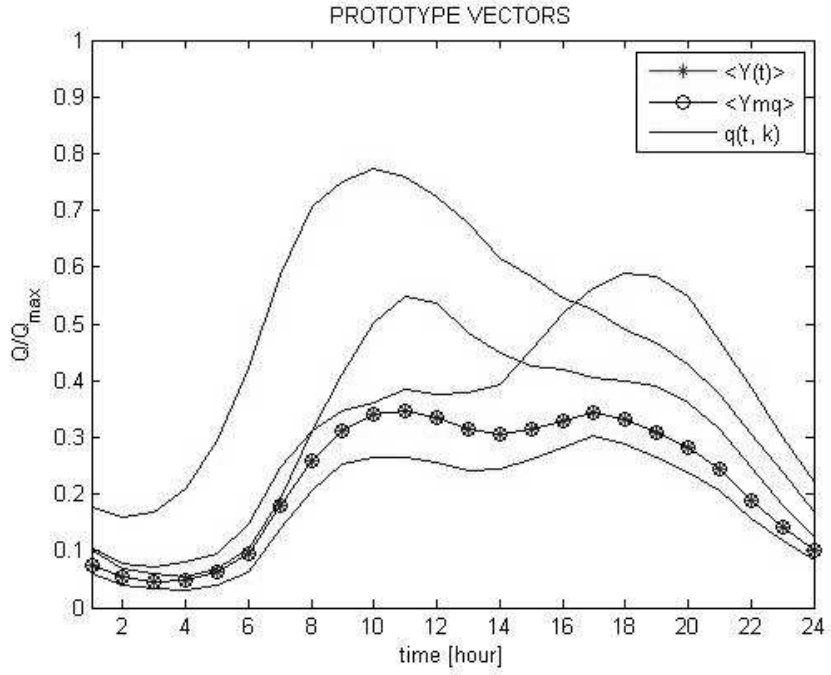

Figure 3: Representative traffic day-vectors formed from traffic flow record presented in Figure 1 (black). The coincidence of the mean vector determined from all day-vectors (-*-) and the mean vector determined from representative vectors (-o-) indicates a proper presentation of the complete phenomenon by the model.

The probabilities corresponding to representative traffic day-vectors are shown in Figure 4 in dependence of representative vector index *k*. It is interesting that 365 day-vectors from the complete year are in this case properly represented by 4 prototype day-vectors that closely correspond to a normal day with the highest *p*, and 3 days around weekends or holidays with approximately 6 times smaller *p*.

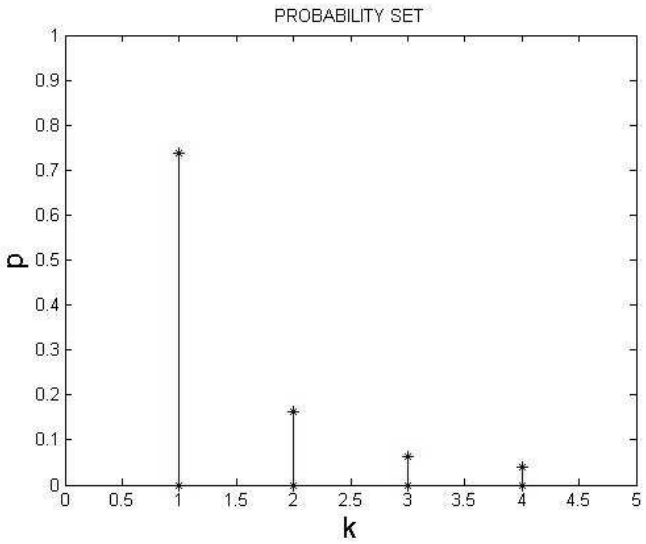

Figure 4: The distribution of probabilities corresponding to representative traffic day-vectors formed from traffic flow record presented in Figure 1.

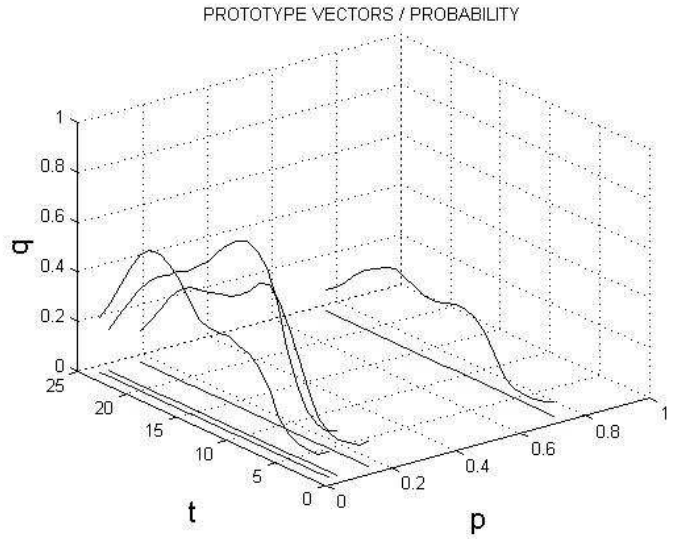

Figure 5: Presentation of traffic day-vectors in dependence of index *t* and probability *p*.

6 Conclusions

We have shown how the PDF of a stochastic variable can be estimated non-parametrically from the experimental data by taking into account the scattering function of the instrument. In order to avoid deficiencies of the kernel estimator we have generalized it to the Gaussian mixture model and introduced the model information, redundancy, and cost function. The most essential terms of the model cost function are the estimation error and the redundancy.

During the minimization of the model cost function the estimation error provides for a proper adaptation of the model to experimental data, while the redundancy prevents excessive effort of experimentation. The search for the model cost function minimum also yields an estimate of a proper number of the acquisition system data storage cells. This number can be surprisingly low in comparison to all acquired data samples since the redundancy and the estimation error are evenhandedly treated in the cost function. The Gaussian mixture model is an effective PDF estimator that is applicable in automatic measurement systems. It could be of especial importance for compression of owerhelming traffic data at their storage in memory media.

The proposed adaptation represents an innovative approach to processing of complex data that could be applied also in modelling of artificial neural networks (Grabec et al., 1997; Haykin, 1999; Leonardis et al., 1998). Adaptation of prototypes by the steps I and II is an essentially non-linear process that resembles self-organized formation of notions in thought processes of intelligent beings and can therefore substitute an intelligent operator at the archiving of traffic data.

Acknowledgement

The author would like to thank the Slovenian Agency for Roads and researchers from EU 7FP Roadidea project for cooperation.

References

[1] Helbing, D. (1997): *Verkehrsdynamik*. Berlin: Springer.

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

[3] Grabec, I. and Sachse, W. (1997): *Synergetics of Measurements, Prediction, and Control*. Berlin: Springer.

[4] Grabec, I. (1990): Self-organization of neurons described by the second maximum-entropy principle. *Biological Cybernetics*, **63**, 403-409.

[5] Kohonen, T. (1989): *Self-Organization and Associative Memory*. Berlin: Springer.

[6] Grabec, I. (2001): Experimental modelling of physical laws. *Eur. Phys. J. B* **22**, 129-135.

[7] Grabec, I. (2005): Extraction of physical laws from joint experimental data. *Eur. Phys. J. B*, **48**, 279-286. (DOI_10.1140/epjb/e2005-00391-0).

[8] Lesurf, J.C.G. (2002): *Information and Measurement*. Bristol: Institute of Physics Publishing.

[9] Parzen, E. (1962): On estimation of probability density function and mode. *Ann. Math. Stat.* **35,** 1065-1076.

[10] Révésh, P. (1991): *Handbook of Statistics*. Krishnaiah, P.R. and Sen, P.K. Eds. Amsterdam: North Holland, **4**, 531-549.

[11] Duda, R.O. and Hart, P.E. (1973): *Pattern Classification and Scene Analysis*, New York: J. Wiley and Sons, Ch.4.

[12] Fukunaga, K. (1972): *Introduction to Statistical Pattern Recognition.* New York: Academic Press, Ch. 6

[13] http://en.wikipedia.org/wiki/Mixture_Model

[14] Cover, T.M. and Thomas, J.A. (1991): *Elements of Information Theory*. New York: John Wiley & Sons.

[15] Kolmogorov, A.N. (1956): To the Shannon theory of information in the continuous signal case. *IEEE Trans. Inf. Theory*, **2,** 102-108.

[16] MacKay, D.J.C. (2003): *Information Theory, Inference, and Learning Algorithms.* Cambridge: Cambridge University Press.

[17] Grabec, I. (2009): Optimal compression of statistical data by minimization of model cost function. *Int. Conf. Applied Statistics*, Ribno (Bled), Slovenia, Sept. 20-23, Program and Abstracts, I-9.

[18] Grabec, I. (2009): Estimation of data redundancy and related statistics. *Metodološki zvezki*, Submitted, Sept. 14,

[19] Haykin, S. (1999): *Neural Networks: A Comprehensive Foundation*. New York: Prentice Hall, Upper Saddle River.

[20] Leonardis, A. and Bischof, H. (1998): An efficient MDL-based construction of RBF networks. *Neural Networks*, **11,** 963-973.