SOME PROPOSALS FOR INCIDENT PREDICTION MODELS IN IN-RESPONSE PROJECT

Lídia Montero and Jaume Barceló
Departament d'Estadística i Investigació Operativa. Edifici U
Universitat Politècnica de Catalunya.

Research involved in European Project E0330
Report Code DR 97/07 (July 1.997)

TABLE OF CONTENTS

1. INCIDENT PREDICTION OVERVIEW

This research is a part of the authors collaboration in the European project IN-RESPONSE (*INcident RESPonse with ON-line innovative Sensing)* in Drive Program. This project is at this point a 2 years project that began in January 1.996. The goal of the project consists on the physical development of a system for Incident Management in urban motorways, that might be available at the local Traffic Control Center. Sites involved in the definition of the tool are: Valencia, Thessaloniki, Munchen, Eindhoven, Paris and Oslo.

The core system is composed by well defined modules:

- Incident Detection: Data collection in real time is the input for a module that develops authomatic incident detection algorithms.
- Incident Prediction: According to actual traffic conditions and previously estimated statistics models, the real time probabilities of incident occurrence have to be computed for short term prevention purposes.
- Incident Verification: Since authomatic incident detection algorithms have a high rate of false alarms, an specific module for verification of the alarms has to be included. This module also deals with civilian calls (by cellular or by signal posts) notifying an incident occurrence. Incident type and severity are set.
- Incident Management: Once an incident has been verified and according to incident type and severity, the incident response units (firemen, ambulances, cleaning units, etc...) have to arrive as soon as possible to the spot.

The authors of this document have focused the research on the study of the state of the art of incident prediction in literature and the selection and development of the best suited proposals for IN-RESPONSE goals: two short-time models for Incident Prediction.

1.1 User needs for incident prediction

The incident prediction model estimates the incident probability on a given stretch of freeway (a road section) in a given period of time (5 or 15 minutes, an hour, a day, a week, a month). Incident prediction models can be used for both preventive (avoiding incidents by improving conditions on the freeway) and curative incident management (reducing the impacts of incidents that do occur). Two different types of models can be distinguished: one that evaluates long-run incident probabilities and one that gives real-time information on incident probabilities. Focusing on real-time information, user groups are defined (see also Table 1-1):

primary users:

⇒ traffic control centres

⇒ regional authorities

These users need real-time information of incident probabilities to determine whether and what sort of traffic management measures need to be activated (like traffic calming or ramp metering). Also, the information can be used to automatically generate warning messages to drivers.

secondary users:

⇒ emergency services

⇒ research institutes

⇒ insurance companies

Secondary users do not always need real-time information, but they can derive the information they need from the real-time information. In general, these will be averages of the real-time probabilities. Researchers will need more detailed information.

Table 1-1 User needs and user groups

user group	user needs	actions
traffic control centre	real-time incident probabilities (for different incident types), expected capacity reduction	traffic calming, ramp metering, estimation of congestion, dissemination of traffic information to drivers
regional authorities	real-time incident probabilities for different incident types, expected capacity reduction; average incident probabilities	traffic management; preventive measures to increase safety, e.g. adapting geometry and traffic demand
emergency services	incident numbers and distribution over the network, for different incident types	optimising the location and deployment of emergency services
researchers	geometry, traffic- and other characteristics of time and location of incident; duration and capacity reduction of incidents	research into the causes and effects of incidents
insurance companies	number of incidents for different types	policy making

1.2 Functional requirements for the incident prediction module

Real-time incident prediction is a new feature, which has not been applied before. An overview of existing, long-term, incident prediction models (see the functional specifications of the incident prediction module in Deliverable 4.1 of the In-Response project) showed that real-time incident prediction differs from existing incident prediction models in an number of ways, most notably in the data requirements. Not all long-term models could accommodate varying circumstances, like traffic or the weather. They incorporated these circumstances by representing them by averages, minima and maxima. Prevailing conditions cannot be represented in such ways.

The Multiproportional Poissonmodel is a model that can be adapted to include variable conditions. It compares the number of incidents occurring under certain circumstances with the amount of time these circumstances prevail, or the length of the road sections where they prevail.

This model will therefore be implemented. Data regarding roadway-, traffic- and other characteristics such as weather and the presence of congestion upstream, is input to the incident prediction module. The output consists of incident probabilities for each of the road sections.

The incident prediction module consists of four parts (see Figure 1.1):

1. data retrieval and preparation module, processing input from monitoring systems and databases
2. incident prediction, the calculation of incident probabilities
3. warning module, for high probabilities
4. traffic management module, supporting decisions to implement traffic management measures (in order to mitigate high incident probabilities resp. to avoid predicted incidents).

The primary user is the operator at a Traffic Management Centre. His goal is to improve conditions on the freeway network, in order to let traffic run smoothly. The output of the incident prediction module is presented to him on screen: the freeway network with incident probabilities indicated by colours for each road section, and (by selecting a road section) an identification of the conditions causing high incident probabilities and suggestions for the application of traffic management measures, taking into account the measures that are already being carried out.

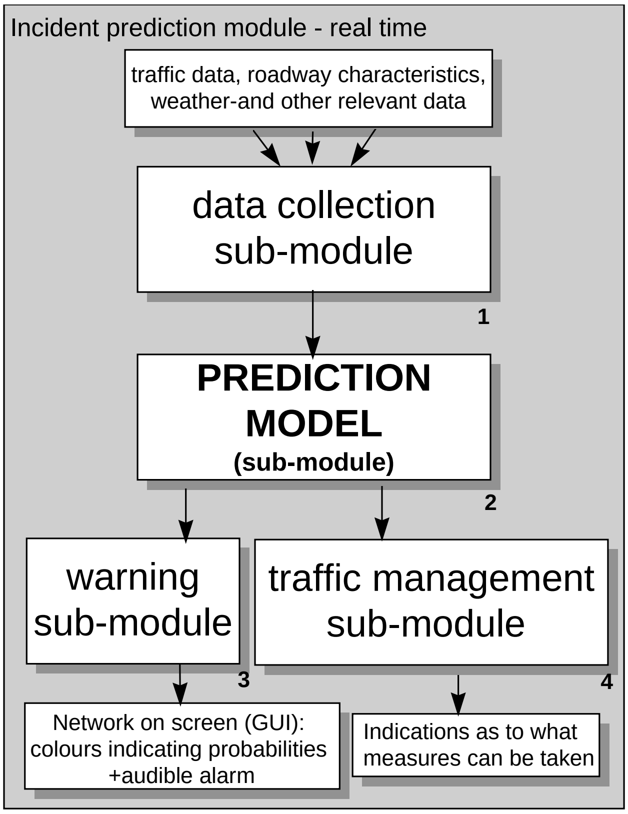
Incident prediction module - real time
traffic data, roadway characteristics, weather-and other relevant data
data collection sub-module
1
PREDICTION MODEL (sub-module)
2
warning sub-module
3
traffic management sub-module
4
Network on screen (GUI): colours indicating probabilities +audible alarm
Indications as to what measures can be taken

Figure 1.1 The incident prediction module

2. INPUT/OUTPUT OF THE INCIDENT PREDICTION MODULE

This is where the actual calculation of the incident probabilities takes place and this module implementation and testing is a direct responsability of the authors in the current IN/RESPONSE project. Per road section and prediction period (5 minutes, ½ hour, or hour, for instance), roadway-, traffic- and other variables are put into the model equation. Per road section, the model gives an estimation of the incident probability. This can be done by using an aggregated model; the Multiproportional Poisson model. This sort of incident prediction models refers to a motorway frame, that is, response and explicative variables for all freeway sections are considered together and the resulting model parameters are valid for all the sections (let us called to this models, *aggregated incident prediction models*).

In this mainframe, geometry characteristics of the section are taken into account and appear to be significant for incident prediction. One might also propose, what we call from this point ahead, the *disaggregated incident prediction models*, that is parameter estimation is going to be performed on a section level, leading to an specific incident prediction model for each freeway section. In disaggregated models, geometry variables are not significant by themselves since they are constant for each section and their effect is reflected in the parameter estimates for traffic and meteo variables; each parameter estimation process is simplified since the number of variables (most of them factors that have to be split in several columns in the design matrix of estimation process) is reduced and the number of observations is also limited to incident occurrence in the current freeway section. In the prediction stage, disaggregated incident prediction models are also easier to apply.

The parameter estimation module for disaggregated models is easier to program, since it deals with less variables and observations, and numerical problems arising frequently in the resolution of the underlying mathematical programs are not going to be so critical.

The main purpose and function of this system module are thus to estimate, short-term incident probabilities and long term expected incidence occurrence in a freeway network that is divided into a number of road sections. The prevailing traffic conditions, meteorological situation and geometric description of each section are input to resulting prediction equations which estimate the short-term probability that an incident occurs.

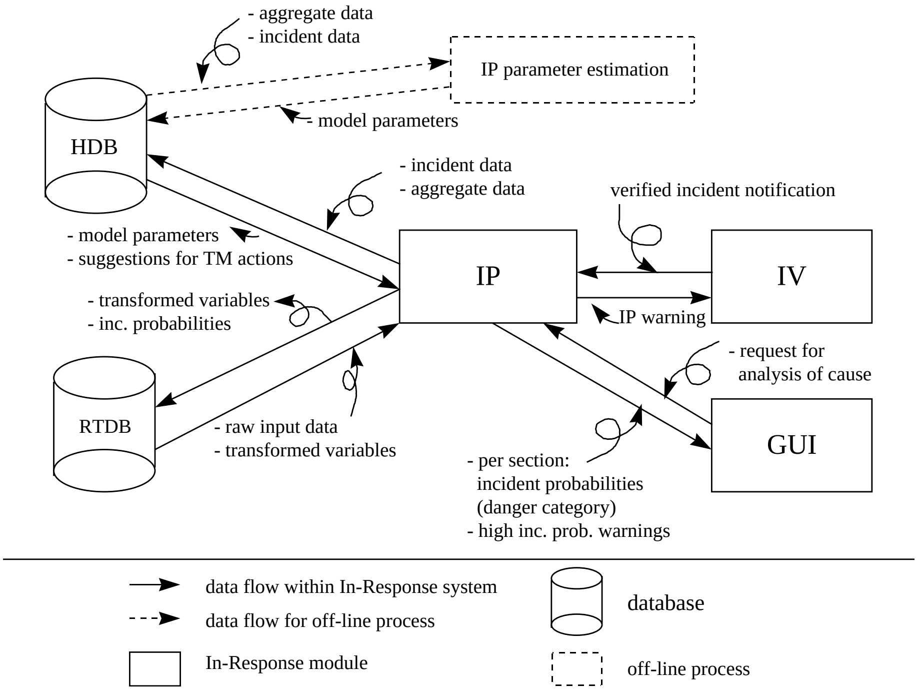

Figure 2.1 Incident Prediction and its environment

Input for the incident prediction module consists for the largest part of data describing the conditions on the freeway network.

2.1 Input Data

A RTDB (Real Time Database) stores all data. This includes the raw data coming in from various monitoring services (e.g. traffic, weather), and roadway data. This data is transformed into variables per road section, suitable for use in the prediction model. Both the raw (updates: for traffic data every 30 seconds, for weather depending on the weather agency providing it) and transformed data (updates every prediction period) are stored in the RTDB.

Apart from this, the model parameters needed for the incident prediction model and the suggestions for traffic management actions are also extracted from the RTDB.

2.1.1 Roadway data

Roadway geometry data can serve as input for both incident prediction and the GUI-module. For incident prediction, roadway data is used to estimate the incident probability. The GUI needs the same sort of information to display the road network on screen.

The proposed roadway geometry variables and their classes are presented in Table 2-1.

Table 2-1 Roadway data

Possible variables:
number of lanes
number of weaving sections in section
number of curves/maximum curve in section
maximum/average grade
number of viaducts in section
presence of lighting
road surface type
number of on/off ramps in section
presence of shoulder lane
locations of emergency phones in section

2.1.2 Traffic characteristics

Traffic data is also used by more than one module (next to IP, incident detection for instance). The information which is interesting for the incident prediction module is presented in the Table 2-2.

Table 2-2 Traffic data

Variable
traffic volume (veh/h)
road capacity (veh/h)
truck percentage (per lane)
speed (distribution) of vehicles (km/h, per lane)
headway distribution
presence of congestion on road section ahead

For some of the variables, it would be desirable to obtain the data per lane, not only using the average value for all lanes together, but also the extreme values, so that the most unfavourable conditions appear.

2.1.3 Weather characteristics

Table 2-3 shows the weather and light data which are considered for inclusion in the model.

Table 2-3 Weather- and light data

Variable
weather conditions (good, rain, snow/sleet, fog, storm)
surface conditions (e.g. dry, wet, slippery)
light conditions/visibility (light, dark, low sun, fog*)

*Fog could be considered a weather condition or a visibility condition

2.1.4 Input for parameter estimation submodule

The parameters of the prediction model, which are used in the incident prediction module, have to be calibrated in advance. For this calibration process, the same sort of input is needed as for the real-time predictions, and a bit more. Apart from data per incident, data describing the conditions on the network over a longer period is also needed.

Some prediction models require explicit quantitative data, real time data as traffic volume or speed variance: as it is the case of the authors' proposal, a Logistic Regression Model for short term incident prediction that includes categorical and continuous explicative variables.

2.2 Output

The prediction model calculates the incident probabilities for the road sections in the network on a given period of time. Incident Prediction for Long-Term Predictions and Short-Term Prediction are different, a detailed description is given in future sections.

3. INCIDENT PREDICTION MODELS

Incident prediction can be done for a long term period or/and a short term period. Each prediction module has the following differences in functions:

- Long term module: The long term module performs one estimation of the number of incidents that can occur during a specific time period (1 day, 1 week, 1 month)
- Short term module: This kind of prevention informs the system about the possibility that an incident will happen in a specific time period (set at 5 minutes).

Table 3-1 Architecture aspects for long- and short term prediction

	Long Term case	**Short Term case**
External components that might receive an **output** from the module.	• Traffic Management • Response module. It can be a criteria for the RU distribution.	• Traffic Management • Warning Alarm • Verification module. The incident have to be verified.
Graphical User Interface -Warning Alarm-	• The system will show, over the map, a shadow of colours with different intensities to represent the number of incidents	• The system shows an ‘incident prediction alarm’ placed. over the map, in the position of the road section with the high incident probabilities
Prediction period	• The integration period is 1 day, 1 week or 1 month.	• The integration period is 5 minutes.

The following submodules can be distinguished in the incident prediction module implementation process:

- Calibration of Parameters : This function should be designed as a generic model, adapted for each site. This is an off-line process, that could be made using a simulator tool, the output of this process will be the mathematical model that will be software programmed in the IP process.
- IP Probability Computation: Once the desired model has been calibration, computation of incident probabilities for a given time interval (short or long term) can be requested.

The Incident Prediction module provides to INRESPONSE system two different options :

- Real time prediction : If an incident is predicted, the system will show an icon located in the map associated section network. If the operator selects the icon, and asks for more information selecting the icon, the system will follow up a little window showing the incident information. (location, value of the incident provability, thresholds of the incident, date and hour of the predicted alarm). This icon will be kept on screen until the alarm is off, or the operator removes it. The operator can confirm or cancel the alarm in a manual form through an option of the general menu.

- Long time prediction (optional): The output of the number of incidents that will occur in a specific time. The output on the screen shows a shadow of colours with different intensities to represent the different number of incidents, over the map on screen.

3.1 Methodology for Incident Prediction

A first consideration concerns the explicative variables involved in incident prediction models. Threre are three groups of variables related to: geometric data, meteorological data and traffic conditions data. According to literature about prediction models: the explicative variables for each data group to be included in a final prediction model heavily depend on the studied motorway and the response variable, and a statistical study should not be avoided before the development of any quasi-automatic module for calibrating (estimation of parameters) incident prediction models.

A methodological scheme is proposed below:

1. Collect data about geometry, meteo and traffic for each road section. All data collected should refer to the same time unit (or the possibility of conversion must consist). All data should also be available for any verified incident.
2. Determine by means of a statistical study that might consider variance analysis or factor analysis, a superset of the significant variables in the current study.
3. Determine for long and short term predictions the significant variables in each data group. A commercial statistical software, as SPLUS, SAS or GLIM, can be used to satisfy this goal. A module for data aggregation to an specific time period should be available for dependent and independent variables in each test site; for example, in a long term prediction base, the number of incidents for a time period of one month or one year should be inferred from collected data.
4. Long and short term explicative variables per group are assumed to be valid for a reasonable period of time and *ad hoc* parameter estimation modules for involved prediction models has to be programmed and included in the IN-RESPONSE environment. Parameter estimation modules should be programmed in such a way that the extension, reduction or substitution of model variables does not affect program code. Re-estimation of model parameters should be performed according to two criteria:

- On a time basis, for example every month for short term prediction models.
- After a certain number of new incidents have occurred since last parameter estimation.

5. The performance of the incident prediction models should be considered in the IN-RESPONSE system, that is, for a given time basis, do all occurred incidents correspond to incident probabilities greater than a given threshold (in a certain confidence level) or are there more high probabilities related to non incident situations than one must expect? This will form the basis for the evaluation of the incident prediction module.
6. After a certain period of time on operation or if predictions are getting worse according to the operator criteria supported by the evaluation of incidence prediction performance module, the explicative variables in prediction models should be reconsidered (point 3).
7. One might hope not to need to modify the underlying modellistic approach, since this would imply the development of a new parameter estimation program, if it was not previously included in the IN-RESPONSE IP parameter estimation module for the test-site.

3.2 Proposals of models for incident prediction

The occurrence of incidents can be analyzed by means of mathematical models. Regression analysis is often used, even linear regression and often a multiplicative model is made linear. The use of multiple linear regression implicitly assumes that the observations results are distributed normally. This assumption is not very realistic since the analysis is specifically concerned with traffic situations in which few incidents occur. The probability that the predicted number of incidents would become negative is not negligible in that case, although the solution to this problem is simple to constrain the prediction to be positive, it has unattractive features: the models may generate unstable estimators and it is preferable the use of a smooth relation between the explicative variables and the expected values of the response variable which leads to the generalized linear model theory. The drawback of an erroneous assumption with respect to the sampling distribution is even greater in the use of the multiplicative model linearised by a logarithmic transformation.

A contribution from Hamerslag deals with the weighted multiproportional Poisson model and illustrates this methods with some applications, the number of incidents is used as the dependent variable, the lengths of road segments where incidents have been observed lead to the introduction of weighted models. Incidents are related to road and traffic characteristics by means of a multiplicative or multiproportional model with the number of incidents per road section assumed to be Poisson distributed. The multiproportional Poisson model employed is based on three assumptions. First, it is assumed that incidents are not correlated and the time interval between two subsequent incidents has an exponential distribution. Second, the expected number of incidents is the product of effects of independent factors weighted by length of the segment. Last assumption relies on the fact that observations from long road segments are more reliable than those from short segments:

Let $\lambda_j = \alpha_J \ell_j$ be the expected number of incidents per road section *j* of type J in a time period of length T, that is the Poisson parameter of the number of incidents distribution. Being α_J , the product of parameter estimates for factor values defining the J class.

The parameters of the incident model are estimated on the basis of the computation of the maximum of the log-likelihood function which is determined by making the value of each first partial derivative (with respect to each parameter) equals to zero, which leads to a set of nonlinear equations with which the coefficient are determined by an iterative method.

The resulting equations are quite similar to those in the trip distribution models in Transportation Models which mostly employ and entropy-maximization approach which may be intuitively interpreted in the following way :

> Consider a system made up of a large number of distinct elements. A full description of such system requires the complete specification of its micro states, as each is distinct and separable. This would involve, for example, identifying each individual incident characteristics. However, for many practical purposes it may sufficient to work on the basis of a more aggregate or meso estate specification, that is in our case, the total number of incidents for all road section with a given characteristic. The basis of the method is to accept that, unless we have information to the contrary, all micro states consistent with our information about meso states are equally likely to occur. It is possible to determine the expression of the number of micro states associated with a given meso state and define a related restricted optimization program which maximizes the *entropy function* and gives as a result the most likely micro states configuration for a set of meso states restrictions.

For a long term prediction model, with dependent variable the number of incidents per road section in a given time period T, almost all involved explicative variables may be considered as factors and in such conditions a multiproportional Poisson model could be calibrated using historic data by a generalized linear model method or by a log-linear model for estimating contingency tables. Time period T to be considered is a month or a year. From a mathematical point of view, data to fit the model are:

1. Dependent variable: number of incidents per road section cross-classification in a time period T for a group of time periods, i.e. if T is a month then a 6 or 12 months should be considered for calibration purposes.
2. Independent variables: for each road section and time period T, a class for each factor involved in the model should be computed. The historic database must contain directly the needed data or a related type of data which allows the classification.
3. Reliability of data is not a problem in long term models since historic data stored in the database are supposed to be verified after being processed for the data collection module. Missing data in any of the variable would cause the elimination of the observation.

From a wider point of view (which may be further investigated in for future incident prediction models), since the Poisson distribution is a type of distribution belonging to the exponential family (as the normal distribution), a generalized linear model with an expected number of incidents per road section type in a time period of length T being the product of factor parameter per road section length may be easily deduced, using as a link function the logarithmic function, the log-likelihood function must be maximized. The first order conditions of this unconstrained non linear program (partial derivatives with respect to parameters equal to zero) define a singular point that in a class of problems we are concerned can be shown to be a maximum, thus the computation of the parameters that maximize the log-likelihood function is equivalent to the solution of a nonlinear system of equations, that requires the definition of an iterative process of linear system resolution, for example, in a Newton-Raphson algorithmic frame.

For real-time prediction models, the above considerations are not easily extended except for the basic IN-RESPONSE prediction model (Multiproportional Poisson Model by Hammerslag), since traffic conditions should be considered in a continuous way, weather conditions are usually critical and time period should be reduced, for example to 5 min. From a conceptual point of view, the dependent variable looks like a Bernoulli type: it is going to occur an incident (1) or not (0) and which is the probability of each value. The calibration of the real-time model should considerate existence or not of incidents for short time periods T' in the immediate days ago for each road section:

1. Dependent variable: existence or not of incidents per road section in a time period T' for a group of time periods, i.e. if T' is an hour then a week should be considered for calibration purposes.
2. Independent variables: for each road section and time period T', a class for each factor involved in the model should be computed and continuous traffic data determined. The historic database must contain directly the needed data or a related type of data which allows the classification or computation by aggregation.
3. Reliability of data is not a problem in the calibration of real time models since historic data stored in the database are supposed to be previously verified.

The Bernoulli distribution is another member of the exponential family and thus the log-likelihood function involved in the calibration of the generalized linear model is going to have a nice formulation and the optimal solution of the unconstrained maximization program is going to be equivalent to the resolution of a nonlinear system of equations. The question now is to define the link function, compute the log-likelihood function and its first order conditions for a maximum.

3.3 State of the Art Practice in Incident Prediction

In the past, several incident prediction models have been developed. They differ in some respects. Most model the relationship between the circumstances arising on road sections and the number of incidents occurring, in a given period of time. There are, however, differences in whether the circumstances are linked to the incident (and, on an aggregate level, to road sections) or to the road section only. Three (groups of) models can be distinguished:

1. Poisson/Negative Binomial regression models
2. idem, but with an extra 'Empirical Bayes' step
3. the Multiproportional Poisson model

The models and their properties are discussed in the following paragraphs.

3.3.1 Poisson/Negative Binomial models

3.3.1.1 Model formulation

The models discussed in this paragraph all use the Poisson distribution to describe the incident occurrence process. In most cases, this leads to the following reasoning. Consider a set of n road sections. Y_i is a stochastic variable that represents the number of incidents on road section i, in a given period of time. y_i is the observed value of Y_i, with yi = 0,1,2,... and i = 1,2,...,n. If the incidents occurring on a road section follow a Poisson distribution, the following equation describes the probability of y_i incidents in period j:

$$P(Y_{ij}) = \frac{e^{-\mu_{ij}} \mu_{ij}{}^{y_{ij}}}{y_{ij}!} \qquad (1)$$

with:

μ_{ij} = the expected number of incidents on road section i in period j.

The model for μ_{ij} is usually written in the following way:

$$\ln \mu_{ij} = X_{ij}\beta \qquad (2)$$

with:

X_{ij} = the vector of network (geometry etc.)- , traffic- and other relevant characteristics, for road section i in period j

ß = vector of coefficients to be estimated

One of the properties of the Poisson is that the variance equals the mean. This property is often violated. In most datasets discussed, the variance is larger than the mean (overdispersion). This is corrected for by adding an gamma-distributed errorterm, thus rewriting equation (2) to a negative binomial model:

$$\ln \mu_{ij} = X_{ij}\beta + \varepsilon_{ij} \tag{3}$$

resulting in the following mean-variance relationship:

$$\mathrm{Var}[Y_{ij}] = E[Y_{ij}][1 + \alpha E[Y_{ij}]] \tag{4}$$

If α is significantly different from zero, the data are overdispersed. If α is equal to zero, the negative binomial reduces to the Poisson distribution. The resulting probability distribution under the negative binomial assumption is:

$$P(Y_{ij}) = \frac{\Gamma(\theta + Y_{ij})}{\Gamma(\theta)Y_{ij}!} u_{ij}^{\theta}(1 - u_{ij})Y_{ij} \tag{5}$$

where:

u_{ij}	=	$\theta(\theta + \mu_{ij})$
θ	=	$1/\alpha$
$\Gamma(.)$	=	a value of the gamma function

Estimating the coefficients can be done using Maximum Likelihood procedures. Using equation (5), the likelihood function for the negative binomial is:

$$L(\mu_{ij}) = \prod_{i=1}^{N} \prod_{j=1}^{T} \frac{\Gamma(\theta + Y_{ij})}{\Gamma(\theta)Y_{ij}!} \left[\frac{\theta}{\theta + Y_{ij}}\right]^{\theta} \left[\frac{Y_{ij}}{\theta + Y_{ij}}\right]^{Y_{ij}} \tag{6}$$

where:

T	=	number of periods measured
N	=	total number of road sections

This function is maximised to obtain the coefficient estimates for α and β.

3.3.1.2 Examples of Poisson- or Negative Binomial models

Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts [Fridstrøm et al., 1995]

For the four Scandinavian countries a study was carried out to compare accident counts and exposure, weather conditions and randomness. The monthly accident count was given per county or province, for several years. A Negative Binomial was used to model the relationship.

The dependent variable was the ***accident count per month per county***. A distinction was made between injury accidents and fatal accidents. The number of road users killed formed a third category. The independent variables considered were:

- gasoline sales (a proxy for exposure; the Danish set used traffic volumes)
- weather conditions (monthly average temperatures, number of days with precipitation/below freezing point)
- the duration of daylight (varying enormously in Nordic countries!)
- changes in legislation and reporting routines
- trend variables/dummy variables for counties and months

Exposure was the most important explaining variable and, after that, weather conditions. Road characteristics were not taken into account in this study.

Estimating truck accident rate and involvement using linear and Poisson regression models [Joshua and Garber, 1990]

This study compared several Linear and Poisson regression models, choosing in the end for Poisson regression. With linear regression, the process could not be adequately described. The aim was to develop a mathematical relationship between the ***number of large truck accidents during a year at a given segment of highway*** and a set of traffic and geomatric variables. Several stretches of highway or interstates were selected. Three environments were distinguished:

1. undivided, four and two lane highways with an AADT[1] of less than 15,000
2. divided, four lane highways with an AADT of less than 15,000
3. divided, four lane highways or Interstates, with an AADT of more than 15,000

For these three environments, models were estimated with the following independent variables:

I. roadway geometry:
   A. number of lanes
   B. lane width
   C. shoulder width
   D. curvature change rate
   E. absolute mean slope
   F. segment length

[1] AADT - Average Annual Daily Traffic

II. traffic variables:

   A. Average Annual Daily Traffic (AADT)
   B. mean speed (all vehicles)
   C. speed variance (all vehicles)
   D. mean speed (trucks)
   E. speed variance (trucks)
   F. mean speed (non-trucks)
   G. speed variance (non-trucks)
   H. percent of large trucks
   I. difference in mean speed between trucks and non-trucks

The models indicated that the slope change rate, the average daily traffic, the percent of trucks and the difference in speed between trucks and non-trucks influence the number of truck involved accidents at a given stretch of highway significantly.

The relationship between truck accidents and geometric design of road sections: Poisson versus Negative Binomial regressions [Miaou, 1994]

In this study, Poisson regression, Negative Binomial regression (NB) and Zero Inflated Poisson regression (ZIP) were compared. The differences found were small, but it was recommended, in the case of overdispersed data (variance greater than mean) to use NB or ZIP regression. Initial relationships can be established using Poisson regression, which is the simplest model of the three.

The dependent variable in the study was the ***number of truck accidents on a road section in one year***. Data were collected for several years. The same road section is considered as several different sections over the years, thus allowing for year-to-year changes in roadway geometry.

Truck exposure (or truck travel, determined by the truck volumes and road section lengths; n_i) is a major variable explaininig truck accident involvement, and is kept out of the x-vector denoting geometric and other characteristics of a section, thus making the equation for the expected number of truck accidents μ:

$$\mu_i = \nu_i \left[e^{\underline{x}_i \underline{\beta}} \right] \tag{7}$$

x is a vector, denoting geometric characteristics, traffic conditions and other relevant attributes:

- dummy variables for the years in which the data were collected
- AADT per lane
- horizontal curvature (in degrees per 100-ft arc)
- length of original horizontal curve

- vertical grade (in percent)
- length of original vertical grade
- deviation of paved inside shoulder width (from ideal width of 12 ft)
- percent trucks in stream
- some interactions

All coefficients found had positive signs, indicating that an increase in the values of these variables results in an increase in the accident frequency, except the coefficient for truck percentage. Apparently, more trucks in the traffic stream results in a lower truck-accident involvement. The possible reason is that (for a constant vehicle density), as truck percentages increase, the frequency of lane changing and overtaking movements by cars decrease. Road users adjust their behaviour.

Effect of roadway geometrics and environmental factors on rural freeway accident frequencies [Shankar et al., 1995]

The effects of roadway geometrics, weather and other seasonal effects on the accident frequencies on rural freeways were studied by Shankar et al.. They estimated models for overall accident frequencies and for specific accident types. This was done with a Negative Binomial model. The NB model was chosen instead of a Poisson model, because the Poisson distribution has the limitation that the mean equals the variance. The dependent variable in the study was ***the accident frequency per month per section***.

The roadsections selected, of a fixed length of 6.1 kilometers, were of a freeway in Washington State. Weather conditions play an important role in the study, as do road geometrics, and the interaction between the two. Four groups of variabels can be distinguished: variables describing horizontal curves, vertical grades, rainfall and sow. The independent variables which were started with were:

- number of horizontal curves (design speed less than 112.6 kph)
- number of horizontal curves (design speed less than 96.5 kph)
- number of horizontal curves (design speed less than 80.5 kph)
- number of horizontal curves in section
- maximum horizontal curve radius in section
- minumum horizontal curve radius in section
- number of vertical curves in section
- maximum grade in section
- minimum grade in section
- average monthly rainfall
- maximum daily rainfall in month
- number of rainy days in the month
- average monthly snowfall
- maximum daily snowfall in month

- number of snowy days in the month

Some of these variables were combined in indicator-type interaction variables (e.g. snowfall-curve interaction indicator: 1 if maximum snowfall greater than 5.1 cm on any given day in the month and at least one curve has a design speed less than 96.5 kph, 0 otherwise), which allows designers to determine thresholds of geometric variables.

The final model chosen included at least one variable of the four categories (curves, grades, rainfall and snow), some section and period indicators and interaction indicators.

Medical conditions, risk exposure, and truck drivers' accidents: an analysis with count data regression models [Dionne et al., 1995]

This study focused on the relationship between medical conditions and traffic safety, following other studies, e.g. studies of the effect of having diabetes on the accident-proneness of drivers. Data on truck drivers, and several variables concerning their medical condition and the circumstances under which they did their work, were collected.

The dependent variable in the study was the ***number of truck accidents per year per driver***. Note that in this case, the unit of measurement did not include road sections. Road characteristics were not taken into account at all, other than type of road.

For each driver in the dataset, the following characteristics were also determined:

- class of driver's permit
- medical condition (such as coronary heart disease, hypertension, diabetis, etc.)
- did the driver own the truck
- distance driven at work
- number of hours behind the wheel
- did the truck pull a trailer
- driving after 8 p.m.
- working radius
- type of road (highway/country road/city street)
- dummy variables indicating observation periods

A Negative Binomial model was chosen, rejecting a Poisson regression model. The results confirmed, in part, earlier findings that diabetic drivers of the permit class for straight trucks have more accidents than drivers in good health. Also, some risk exposure variables were significant.

Incident Management focuses primarily on measures affecting the environment in which road users perform their driving tasks. Therefore, medical and other conditions affecting driving skills, are left out for the time being. In the future however, this could certainly be part of preventive Incident Management.

3.3.2 Poisson/Negative Binomial regression with Empirical Bayes Estimation [Persaud and Mucsi, 1995]

The purpose of this study, **"Microscopic accident prediction models for two-lane rural roads",** was to estimate the accident potential of road sections, based on traffic counts and geometric characteristics. This was done for several incident types (single/multi/all vehicle). The dependent variable was the ***expected number of accidents during T hours on a section of L km*** (E(m) (µ in previous paragraphs)). The fundamental estimator for E(m) is given by equation 8:

$$E(m) = aLTF^{b} \tag{8}$$

with:

L = length of road section

T = length of time period

F = traffic volume

a,b = parameters to be estimated in a regression model

This model was estimated with a Generalized Linear Modelling technique, allowing the specification of different error terms. The Negative Binomial distribution was considered more appropriate than the Poisson or Normal distribution. The variance is then related to E(m) as follows:

$$Var(m) = E(m)^{2} \; / \; k \tag{9}$$

where k can be estimated using a maximum likelihood procedure that assumes that each squared residual of the regression model is an estimate of Var(m) and that each count comes from a Negative Binomial distribution with mean E(m) and variance given by equation 9.

Accident counts are usually small, and the variance relatively large. The value of the expected number of accidents for a specific site is, because of that, often not very useful. For this reason, the model also includes an emperical Bayesian (EB) procedure. This procedure combines the regression estimate E(m) and the short term (observed) accident count (x) of the specific site.

The empirical Bayesian estimate of accident potential is then:

$$E(m|x) = wE(m) + (1 - w)x \tag{10}$$

where

$$w = [1 + Var(m) / E(m)]^{-1} = [1 + e(m) / k]^{-1} \tag{11}$$

The estimation of the variation in $(m \mid x)$ can be estimated by:

$$Var(m|x) = (x + k) / [1 + (k / E(m)]^{-1} \tag{12}$$

The geometric characteristics taken into account were lane- and shoulder width. Different models were estimated for all combinations possible (narrow lanes, wide shoulders, wide lanes, wide shoulders, etc.).

So, sections with different geometric characteristics were not put into one model together, but were put into different models. If more geometric characteristics were taken into account, this would mean an increase in the number of models which have to be estimated.

3.3.3 The Multiproportional Poisson model [Hamerslag et al., 1982]

The model developed here (**"Analysis of Accidents in Traffic Situations by Means of Multiproportional Weighted Poisson Model"**) describes how the expected number of accidents depends on road and traffic characteristics. The model has a multiplicative form and ***the expected number of accidents on a road section in a given period of time*** (the dependent variable) is a funtion of the estimated parameters and the length of the section studied.

The independent variables are all divided into classes; the number of factors and of their classes determines the number of coefficients to be estimated.

The model equation is the following:

$$\mu_{klm} = E(Y_{klm,L}) = \beta_{1k} \cdot \beta_{2l} \cdot \beta_{3m} \cdot L_{klm}$$

where:

m_{klm} = expected number of incidents on a section with characteristics $ß_{1k}$, $ß_{2l}$ and $ß_{3m}$ and length L;

L_{klm} = total length of the road section that belongs to the categories k,l and m;

$ß_{1k}$...$ß_{3m}$ = the coefficients (representing the different infuential factors $ß_1$, $ß_2$ and $ß_3$);

k..m = classes (each factor is divided into several classes; no continuous variables).

The probability of y_{klmn} incidents is:

$$P(y_{klmn}) = \frac{e^{-\mu_{klmn}} \mu^{y_{klmn}}}{y_{klmn}!} \tag{14}$$

To estimate the coefficients a, b, c and d, the log-likelihood finction L^* is maximized:

$$L^* = \sum\sum\sum\sum \ln P(y_{klmn}) \tag{15}$$

The maximal value of the log-likelihood can be found by setting the partial derivatives to zero:

$$\frac{\partial L^*}{\partial \hat{\beta}_{1k}} = \sum_{l} \sum_{m} (-\beta_{21}\beta_{3m} \ldots L_{klm}) + \sum_{l} \sum_{m} (\frac{y_{klm}}{\hat{\beta}_{1k}}) = 0, \forall k \tag{16a}$$

and also

$$\frac{\partial L^*}{\partial \hat{\beta}_{21}} = 0, \forall l \tag{16b}$$

$$\frac{\partial L^*}{\partial \hat{\beta}_{3m}} = 0, \forall m \tag{16c}$$

The coefficients b_1, b_2 and b_3 can be determined by solving the following set of (non-)linear equations:

$$\hat{\beta}_{1k} = \frac{y_{k..}}{\sum_{k} \sum_{l} \sum_{m} (\hat{\beta}_{21}\hat{\beta}_{3m} \ldots L_{klm})}, \forall k \tag{17a}$$

$$\hat{\beta}_{21} = \frac{y_{.m.}}{\sum_{k} \sum_{l} \sum_{m} (\hat{\beta}_{1k}\hat{\beta}_{3m} \ldots L_{klm})}, \forall l \tag{17b}$$

$$\hat{\beta}_{3m} = \frac{y_{..m}}{\sum_{k} \sum_{l} \sum_{m} (\hat{\beta}_{1k}\hat{\beta}_{21} \ldots L_{klm})}, \forall m \tag{17c}$$

with

$$\sum_{l} \sum_{m} y_{klm} = y_{k..}, \forall k \tag{18a}$$

$$\sum_{k} \sum_{m} y_{klm} = y_{.l.}, \forall l \tag{18b}$$

$$\sum_{k} \sum_{l} y_{klm} = y_{..m}, \forall m \tag{18c}$$

The roadway characteristics studied include:

- Average Daily traffic (for motor vehicles and bicycles)
- truck percentage
- lane width, shoulder width, bicycle lane width, median width
- horizontal curves
- type of obstacle and obstacle distance
- permitted speed
- access points
- sight distance

The model results are well in line with expectations; high traffic volumes result in more accidents, as do the presence of narrow lanes, obstacles etc.

3.3.4 Conclusions on the Current Practice

The aim of an incident prediction model is to give an estimation of the accident frequency on a given stretch of road in a given period. The model can be used in several ways:

1. To give long-term average incident frequencies, to be used in safety analysis: which factors influence the number of incidents and can these factors be influenced to improve safety?
2. To give short-term expected incident frequencies (real-time), to be used in a traffic management system: what are the current incident probabilities, should measures be taken *now* to bring the incident probabilities down, and what measures could these be?
3. To provide incident probabilities for models, simulating the incident recovery process. The benefits of different Incident Managament measures can be evaluated. A simulation model can

compare several possible measures and give indications as to what measures would have most impact.

All model types discussed in the previous paragraphs were long-term incident prediction models. Data of incidents that occurred over a certain period of time were collected, and the influence of different variables was investigated. The aim in these studies was to establish which factors influence accident frequencies. The results can be used to work out ways to prevent incidents from occurring.

In some cases, however, it is a temporary combination of factors causing high incident probabilities. An example of such a combination is adverse weather, combined with medium to high flow rates. Though medium to high volumes alone would not always result in dangerous situations, the combination with adverse weather could mean that accident probabilities pass a threshold, indicating that measures should be taken *at this moment*.

For Incident Management, it would be an interesting feature if traffic control operators can determine whether measures should be taken, considering the prevailing incident probabilities.The question then arises, which combinations are potentionally dangerous? How can varying circumstances best be modelled? A real-time incident management should be able to address this question.

The distinction between long-term and short-term incident frequencies leads to the distinction between two types of measures to bring down the number of incidents:

1. permanent changes, in roadway geometry or network lay-out or the improvement of pavement conditions. Long-term average incident frequencies can point out which characteristics influence traffic safety negatively, and lasting measures can be taken. These measures are therefore mostly infrastructural ones.
2. short term changes, following Dynamic Traffic Management measures (information dissemination, traffic calming, ramp metering, lowering speed limits). Prevailing conditions are constantly monitored and, when a dangerous situation occurs, measures are taken to influence those variables or combinations of variables, which cause incident probabilities to be high. Dynamic Traffic Management aims mostly at influencing traffic characteristics. It can also be used to raise the level of attention of road users, using radio messages or variable message signs.

Weather conditions cannot be influenced. They can be monitored, and their influence on incident probabilities estimated. Measures to alleviate the negative impact of adverse weather can aim on improving long-term, static conditions or short-term (traffic) characteristics.

This study focuses on the real-time application of an incident probability model. The next paragraphs therefore discuss the advantages and disadvantages of the discussed modeltypes, with the emphasis on the question how they can be used for real-time incident prediction.

3.3.4.1 Advantages and disadvantages of the revised model types

Any short-term incident probability model can also be used to generate long-term incident probabilities, but the other way round things are more complicated. This has to do with the way in which variables, influencing the number of incidents, are incorporated in the model. For long-term predictions, it is sufficient to work with average values, minima and maxima, or to use variables like 'the number of rainy days per month'. These variables might give a very good result in long-term models, but cannot be used for short-term predictions. Prevailing conditions cannot be represented by averages, minima or maxima. If it is raining, the input for an incident prediction model should be that it is raining, and not what the average rainfall this month is. A single shower does not give any indication of what the average rainfall this month is going to be.

The Poisson/Negative Binomial models (with or without Bayesian techniques applied) and the Multipropotional Poissonmodel have a fundamentally different approach, each with their own limitations. They are discussed in the next two paragraphs. A Bayesian approach can be used in all models; the idea behind it stays the same as in the model described in paragraph 2. The model described there is not suitable when many variables are investigated, so it is left out of the discussion here.

3.3.4.1.1 Poisson/Negative Binomial models (NB)

Poisson and Negative Binomial-models can handle both discrete and continuous variables. A variable can thus have any value which is measured (traffic volumes in vehicles per hour or day, number of curves in a section, maximum rainfall, etc), using all information available. Also, dichotomous variables, like the presence of a weaving section, can be included (value: 0-not present, or 1-present).

As mentioned before, variable circumstances are difficult to deal with in these models. They are usually represented by averages, minima and maxima. This makes, on the other hand, collection of the data rather simple. Counting the number of curves in a section, or the number of rainy days is easier than measuring the exact length of curves, or the time that it rains. It also means that the circumstances at the time that the incident took place, do not have to be measured. It suffies to know the average and extreme circumstances on a road section, and how many incidents occurred, on that same road section, during a certain period of time. Table 1 gives an example of the structure of the inputdata.

Example of input data for NB-models

Road section	Variable 1: # of curves	Variable 2: minimum radius (m)	Variable 3: # of rainy days	Variable 4: maximum rainfall per month (mm)	Dependent variable: # of incidents per month
1	3	2000	5	40	4
2	0	-	5	40	2
3	1	5000	4	35	0
|	|	|	|	|	|
4	2	4500	3	20	1

3.3.4.1.2 Multiproportional Poissonmodel (MP)

The multiproportional Poisson model can incorporate varying circumstances like weather, or traffic volumes. All variables have to be divided into classes, however, so some information is bound to be lost.

The data needed is the following:

- under which circumstances incidents occur
- on what percentage of the total selected road length or during what percentage of the time these circumstances are present

So, for each incident it has to be known on what section it occurred (determining the value of the static variables), and what the prevailing conditions were: what was the weather like, what traffic volumes were measured, etc.. If the incident data contain a variable indicating at what time the incident occurred, most of the data can be collected.

A problem arising here is how the variable data is aggregated. Is it sufficient to aggregate to hourly values, or 5-minute values, and how well is recorded at what time the incident occurred?

Apart from all circumstances *during the incident*, it also has to be known what percentage of the time (or on what percentage of the length of the road) these circumstances are prevailing. Therefore, you also need data of times when no incidents occurred. The multiproportional Poissonmodel needs the maxmimum amount of information available, and the question is whether it is feasible to collect and process all necessary information real-time. Table 2 gives an example of the structure of the inputdata for the multiproportional Poissonmodel.

Example of inputdata for MP-models

Variables	% of length/ % of time	# of incidents
Number of curves:		
1. 0	70	67
2. 1-2	20	3
3. >3	10	10
Minimum curve radius:		
1. 0-1000m	2	6
2. 1000-3000m	15	10
3. 3000-6000m	13	9
4. >6000	70	55
weather condition:		
1. good (dry)	85	50
2. rainy	12	24
3. storm	1	2
4. ice, snow	1	3
5. fog	1	1
traffic volume:		

1. 0-1000 veh/h 2. 1000-2500 veh/h 3. 2500-5000 veh/h 4. 5000-6500 veh/h 5. >6500 veh/h	50 15 15 12 8	25 11 14 18 12

3.3.5 References

- Dionne, G., D. Desjardins, C. Laberge-Nadeau and U. Maag (1995), “Medical conditions, risk exposure, and truck drivers’ accidents: an analysis with count data regression models”, Accident Analysis and Prevention, Vol.27, No.3, pp.295-305, 1995.

- Fokkema, H.J. (1987), Weersgesteldheid en verkeersveiligheid”, Traffic Test, 1987.

- Fridstrøm, L., J. Ifver, S. Ingebrigtsen, R. Kulmala and L Krogsgard Thomsen (1995), “Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts”, Accident Analysis and Prevention, Vol.27, No.1, pp.1-20, 1995.

- Hamerslag, R., J.P. Roos and M. Kwakernaak (1982), "Analysis of Accidents in Traffic Situations by Means of Multiproportional Weighted Poisson Model", Transportation Research Record 847, 1982.

- Joshua, S.C. and N.J. Garber (1990), “Estimating truck accident rate and involvement using linear and Poisson regression models”, Transportation planning and technology, Vol.15, pp.41-58, 1990.

- Koornstra, M.J. (1992), “Verkeersonveiligheid bij mist”, Leidschendam, SWOV, 1992.

- Miaou, S-P. and H. Lum (1992), "Statistical evaluation of the effects of highway geometric design on truck accident involvements", Transportation Research Record 1407, Washington D.C., 1992.

- Miaou, S-P. (1994), "The relationship between truck accidents and geometric design of road sections: Poisson versus Negative Binomial regressions", preprint, 73rd Annual Meeting, TRB, Washington D.C., januari 1994.

- Persaud, B.N. (1992), “estimating accident potential of Ontario road sections”, Transportation Research Record 1327, TRB, Washington D.C., 1992 (??).

- Persaud, B.N. and K. Mucsi (1995), “Microscopic accident prediction models for two-lane rural roads”, preprint, 74th Annual Meeting, TRB, Washington D.C., January 1995.

- Shankar, V., F. Mannering and W. Barfield (1995), “Effect of roadway geometrics and environmental factors on rural freeway accident frequencies”, Accident Analysis and Prevention, Vol. 27, No. 3, pp. 371-389, 1995.

4. CALIBRATION OF MODEL PARAMETERS

4.1 Survey of Statistical Estimation Theory

For several decades linear models of the form

$$Y = X\beta + \varepsilon$$

in which the elements representing the errors are assumed to be independent and identically normal distributed, have formed the basis for most analyses of continuous data. X are called independent or explanatory variables, maybe continuous or categorical (this last situation implies the inclusion of dummy variables to the design matrix). Y is the dependent or response variable.

Recent advanced in statistical theory and computer software allow to use methods analogous to those developed for linear models in the following situations:

- The response variables have distributions other than the Normal distribution, they may even be categorical rather than continuous.
- The relationship between the response and explanatory variables need not be of the simple linear form.

One of these advances has been the recognition that many of the nice properties of the Normal distribution are shared by a wider class of distributions called the *exponential family of distributions*. A second advance is the extension of the numerical methods for estimating parameters, from linear combinations like $X\boldsymbol{\beta}$ to functions of linear combinations $g(X\boldsymbol{\beta})$. In theory, the estimation procedures are straightforward, but in practice they involve a considerable amount of computation so that they have only become feasible with the development of computer programs for numerical optimization on nonlinear functions, that are included in many statistical packages.

4.1.1 Exponential Family of Distributions

If we consider a single random variable Y whose probability distribution function, if it is discrete, or probability density function, if it is continuous, depends on a single parameter θ, then the distribution belongs to the exponential family if it can be written in the form

$$f(Y;\boldsymbol{\theta}) = s(Y)\,t(\boldsymbol{\theta})\,\mathrm{e}^{a(Y)b(Y)}$$

If $a(Y) = Y$, the former distribution is said to be in the canonical form.

Many well-known distributions belong to the exponential family. For example, Poisson, Binomial and Bernoulli distributions, all of them appearing in incident prediction models to be considered for IN-RESPONSE implementation, can all be written in the canonical form.

For example, the probability function for the a discrete Poisson random variable Y is

$$f(y;\boldsymbol{\lambda}) = \frac{\boldsymbol{\lambda}^y e^{-\lambda}}{y!} = \exp(y\log\boldsymbol{\lambda} - \boldsymbol{\lambda} - \log y!)$$

which is in the canonical form.

Let L be the log-likelihood function and U the first derivate of L with respect to θ, called the score. Then for any distribution the following properties can be shown to hold:

$$\mathrm{E}(U) = 0$$

and

$$\mathrm{var}(U) = \mathrm{E}(U^2) \quad \textit{Information matrix}$$

And thus, the log-likelihood function for distributions belonging to the exponential family can be expressed as:

$$L = \log f = a(y)b(\theta) + c(\theta) + d(y)$$

4.1.2 Generalized Linear Models

The idea of a generalized linear model is defined in terms of a set of independent random variables $Y_1, \dots, Y_N$ each with a distribution from the exponential family with the following properties:

- The distribution of each variable Y depends on a single parameter θ_i and the distributions of all the Y variables are of the same form (all Normal or all Binomial), thus the log-likelihood function of the joint probability density is ,

$$L = \log f = \sum_{i=1}^{N} y_i\, b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)$$

For model specification, the parameters θ_i are usually not of direct interest (since they may be one for each observation). For a generalized linear model we consider a smaller set of parameters $\beta_1, \ldots, \beta_p$ *(p<N)* such that a linear combination of β´s is equal to some function of the expected value μ_i of Y_i, that is $g(\mu_i) = X_i^T \beta$, where *g* is monotone, differentiable and is called the *link function*.

β is a vector of parameters and X´s is the design matrix with column vectors equal to the explanatory variables, either covariates or dummy variables for levels of factors.

The method of maximum likelihood is used for statistical estimation of generalized linear model parameters. Usually the estimates have to be obtained numerically by an iterative procedure which turns out to be closely related to weighted least squares estimation.

Estimators for **β** (denoted for **b**) are often obtained by differentiating the log likelihood function with respect to each element of β and solving the simultaneous system of (nonlinear) equations:

$$\frac{\partial\, L(\boldsymbol{\beta};\, \vec{y})}{\partial\, \boldsymbol{\beta}_j} = 0 \qquad j = 1, \ldots, p$$

It is necessary to check that the solution do correspond to maxima of L function, by verifying that the matrix of second derivatives evaluated at the singular point is negative definite, and also if there are any values at the edges of the parameter space value set which give local maxima of L. For all the models considered in IN-RESPONSE Incident Prediction module, there is only one maximum and it corresponds to the solution of the former system of equations.

An important property of maximum likelihood estimator is that for any link function of the parameters, the maximum likelihood estimator of g(β) is g(b), which it is called the *invariance property* of maximum likelihood estimators. A consequence is that we can work with any link function of the parameters which is convenient for maximum likelihood estimation (simplifies the system of equations expression) and then use the invariance property to obtain maximum likelihood estimates for the required parameters. Other properties of maximum likelihood estimators include consistency, sufficiency and asymptotic efficiency.

We wish to obtain the maximum likelihood estimators of the parameters **β** for the generalized linear models defined previously. The loglikelihood function for independent responses $Y_1, \dots, Y_N$ is

$$L(\theta; \vec{y}) = \sum_{i=1}^{N} y_i \, b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)$$

where

$$\mathrm{E}(Y_i) = \mu_i = {}^{-c'(\theta_i)}\!\big/_{b'(\theta_i)}$$

and

$g(\mu_i) = X_i^T \beta = \eta_i$ where g is some monotone and differentiable function.

A property of the exponential family distributions is that they satisfy enough regularity conditions to ensure that the global maximum of the log-likelihood function is given uniquely by the solution of the equations leading to the first optimality conditions , this is

$$\frac{\partial \, L(\boldsymbol{\theta}_i ; \vec{y})}{\partial \, \boldsymbol{\theta}_i} = 0 \quad i = 1, \dots, N \quad \text{or} \quad \frac{\partial \, L(\beta ; \vec{y})}{\partial \, \beta_j} = 0 \quad j = 1, \dots, p$$

and it can be shown that

$$\frac{\partial \, L(\boldsymbol{\beta}; \vec{y})}{\partial \, \boldsymbol{\beta}_j} = 0 = U_j = \sum_{i=1}^{N} \frac{(y_i - \boldsymbol{\mu}_i) x_{i\,j}}{\mathrm{var}(Y_i)} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i} \right) \quad j = 1, \dots, p$$

where $x_{i\,j}$ *is jth element of* X_i^T *for* $j = 1, \dots, p$

In general, the former equations are non linear and they have to be solved by numerical iteration. If the *Newton-Raphson method* is applied then the *m*th approximation is given by

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} - \left[\frac{\partial^2 \, L}{\partial \beta_j \, \partial \beta_k} \right]^{-1}_{\beta = \mathbf{b}^{(\mathbf{m-1})}} \left[\frac{\partial L}{\partial \beta_j} \right]_{\beta = \mathbf{b}^{(\mathbf{m-1})}}$$

where $\left[\dfrac{\partial^2 \, L}{\partial \beta_j \, \partial \beta_k} \right]^{-1}_{\beta = \mathbf{b}^{(\mathbf{m-1})}}$

is the matrix of second derivatives of L evaluated at $\beta = \mathbf{b}^{(\mathbf{m-1})}$.

An alternative procedure which is sometimes simpler than the Newton-Raphson method is called the *method of scoring*. It involves replacing the matrix of second derivatives by the matrix of expected values

$$\mathrm{E}\left[\frac{\partial^2 L}{\partial\beta_j\,\partial\beta_k}\right]$$

that can be shown to be equal to the negative of the variance-covariance matrix of the $U_j{}'s$ and minus the information matrix $\Im_{jk} = \mathrm{E}\left[U_j U_k\right] = \sum_{i=1}^{N} \frac{x_{ij}\,x_{ik}}{\mathrm{var}(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$ and thus

$\Im$ can be written as $\Im = \mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X}$ where W is a NxN diagonal matrix with elements $w_{ii} = \sum_{i=1}^{N} \frac{1}{\mathrm{var}(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$ and hence the iterative equation for method of scoring can be written as

$\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X}\mathbf{b}^{(\mathbf{m})} = \mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{z}$ with the element of **z**

$$z_i = \sum_k x_{ik}\,b_k^{(m)} + \left(y_i - \mu_i\right)\left(\frac{\partial\eta_i}{\partial\mu_i}\right) \quad i = 1,\ldots,N$$

This has the same form as the normal equations for a linear model obtained by weighted least squares, except that it is to be solved iteratively because in general **z** and **W** depend on **b**. Thus for generalized linear models maximum likelihood estimators are obtained by an *iterative weighted least squares* procedure.

Normal equations for a typical least squares method (observations identically distributed with the same variance):

$$\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{b} = \mathbf{X}^{\mathbf{T}}\mathbf{y}$$

Normal equations for a weighted least squares method with **V** defined as the variance-covariance matrix of the observed variables:

$$\mathbf{X^T V^{-1} X b} = \mathbf{X^T V^{-1} y}$$

4.2 Short-Term Prediction: Estimation of parameters

This proposal from UPC is a particular case of a generalized linear model, where the dependent variables Y´s are assumed to be Bernoulli distributed and according to the experience with these models a logit link function may be suitable. Let us readapt in the following the general theory of generalized linear models to the present particular situation.

Let index *k* indicate the time period observation.

Let y_i indicate an observation of a road section *j* in a time period *k* that belongs to type J.

Let $\Pi_i = 1 \big/ 1+e^{-X_i^T\beta}$ *and link function* $\ln\left(\frac{\Pi_i}{1-\Pi_i}\right)$ be the expected number of incidents (probability) per road section j of type J in a time period of length T’, that is the Bernoulli parameter of the existence of incidents distribution and logit link function.

Let $\boldsymbol{\eta}_i = X_i^T\boldsymbol{\beta}$ be the link function value for observation *i* expressed in vectorial form.

The log-likehood function for the logit link is defined as:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i\, \mathbf{X_i^T}\boldsymbol{\beta} - \sum_{i=1}^{N}\log\left(1+e^{\mathbf{X_i^T}\beta}\right)$$

The method of scoring iterates a process (on *m*) for the computation of the singular point of the log-likelihood function is defined below. The algorithmic scheme for estimating logit-model equations by the scoring method requires numerical solution since **W** and **z** (defined in the general formulation) are non linear functions of **b**:

1. Start with estimates $\mathbf{b}^{(0)}$. One particular choice of initial estimates is $\mathbf{b}^{(0)} = 0$.
2. At iteration *m+1*, compute the new estimates by solving:

$$\begin{array}{lcl} \mathbf{b^{(m+1)}} & = & \mathbf{b^{(m)}} + \left(\mathbf{X^T V^{(m)} X}\right)^{-1}\mathbf{X^T}\left(\mathbf{y} - \mathbf{p^{(m)}}\right) \\ \textit{where } \hat{y}_i = p_i^{(m)} & = & 1 \big/ 1+e^{-\mathbf{X_i^T b^{(m)}}} \\ \textit{and } \mathbf{V^{(m)}} & = & diag\left\{ p_i^{(m)} \big/ 1-p_i^{(m)} \right\} \end{array}$$

3. Iterations continue until $\mathbf{b^{(m+1)}} \approx \mathbf{b^{(m)}}$, according to a prefixed tolerance.

Notice that when convergence takes place $\left(\mathbf{X^T V^{(m)} X}\right)^{-1}\mathbf{X^T}\left(\mathbf{y}-\mathbf{p^{(m)}}\right) \approx \mathbf{0}$

and thus the estimating equations are approximately satisfied $\mathbf{X^T y} = \mathbf{X^T}\,\mathbf{p}^{(\mathbf{m})} = \mathbf{X^T \hat{y}}$. Conversely, if $\mathbf{X^T y}$ is very different from $\mathbf{X^T}\,\mathbf{p}^{(\mathbf{m})} = \mathbf{X^T \hat{y}}$ there will be a large adjustment in **b** from one iteration to the next.

The log-likelihood is then

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i\, \mathbf{X_i^T}\boldsymbol{\beta} - \sum_{i=1}^{N} e^{\mathbf{X_i^T}\boldsymbol{\beta}}$$

A dummy variable for road section observation in different time intervals should be included in the model (in design matrix X). Independent variables considered in the model may be covariates or factors. In the case of covariates, they are directly represented in the design matrix X. In the case of factors, each factor value should be splitted into a dummy variable to be included in the design matrix, but finally, it has to be reduced (by transformation) to avoid singularities.

The technique of including dummy variables permits to enter qualitative independent variables into a regression equation (generalized or nor) and model interactions between qualitative and quantitative variables in a regression. A three category factor classification may be entered into the regression equation by coding two dummy variables: D1 and D2.

D1 has 1 value for observations of the factor in category 1 and 0 for other categories.

D2 has 1 value for observations of the factor in category 2 and 0 for other categories.

Observations in the third category of the factor are coded 0 for both dummy variables and are usually called the baseline category with which the other groups are compared. If we are interested in testing the null hypothesis of no effect of factor values, it can be done by the incremental sum of the squares approach in normal regression models.

In general, for a polychotomous independent variable with m categories, we need to code m-1 dummy regressors, so that Dj=1 when an observation falls in category j, and Dj=0 otherwise; and consequently, all Dj=0 for and observation in the category m.When there is more than one qualitative independent variable, and if we assume that these variables have additive effects, we simply code a set of dummy regressors for each one. To test the hypothesis that the effects of a qualitative variable are nil, we delete its dummy regressor from the model and compute the incremental sum of squares.

Two independent variables are said to interact in determining a dependent variable when the effect of one depends upon the value of the other. The linear additive models specify the absence of interactions. The dummy variable regression model may be modified to accommodate interactions between quantitative and qualitative independent variables. Since the two concepts are frequently confused, we must take into account that interaction and correlation of independent variables are logically and empirically distinct: two independent variables may interact wether or not they are related to one another statistically. In simple normal regression models, additive dummy variable regression assumes parallel regression lines across the several categories of a qualitative variable. If these regression lines are not parallel, then the qualitative variable interacts with one or more quantitative independent variables. The dummy regression models may be modified to reflect this interactions in a simple way: define one interaction dummy variable, $I_{ij} = D_j X_i$, for each dummy variable Dj by including the a new column in the design matrix being the product of Dj by X_i. To test the interaction effect, it is only a question of computing the increment in the sum of squares when the interaction dummy variables are not included in the model; in generalized linear models, a likelihood-ratio chi-square test may be employed for contrasting two models, when one model is a restricted version of the other:

Let L0 be the maximized likelihood for a model that sets the *k (k<p)* coefficient to zero, and L1 the maximized likelihood for the complete model (*p* parameters), then the likelihood ratio test statistic is defined as:

$$G_0^2 = -2\log\frac{L0}{L1} = 2(\log L1 - \log L0) \quad \approx \chi^2_{\nu=p-k}$$

Test of this form are analogous to incremental sum of squares F-tests for normal linear models.

The mathematical model and statistical process has been described now, but the IN-RESPONSE parameter estimation module has to be include a submodel of data preprocessing, which creates the proper design matrix according to the dummy variable model described before for the estimation process. Model test is not a feature to be included in the estimation module: the significant variables for each test-site are assumed to be known (determined by a previous model selection phase) and it only deals with computation of model parameters. For simplicity in the interpretation of the model parameters, interactions between factors and covariates are considered only for the simplest situation: one factor per one covariate.

Input to IP PREPROCESS

Number of observations;

Dependent variable values;

```
for each covariate
        Covariate identifier
        Independent quantitative variable values (covariate values)
endfor;

for each factor
        Number of categories
        Independent factor values for each observation
endfor;

for each considered qualitative interaction
        Interaction identifier, dimension of the interaction;
        Identifier of Involved Factor 1, Identifier of Involved Factor 2, ...
endfor;

for each considered qualitative-quantitative interaction
        Interaction identifier;
        Identifier of Factor;
        Identifier of Covariate
endfor

end Preprocess
Output: Design Matrix X.
```

4.3 Long-Term Estimation Model: Incidents Poisson distributed.

For a long term prediction model, with dependent variable the number of incidents per road section for a given time period T, all involved explicative variables may be considered as factors and in such conditions a multiproportional Poisson model could be calibrated using historic data for:

- The expected frequencies for each cross-category (special case where all independendent variables are qualitative, except one that acts as a weight and may be taken into account as initial value for expected cross-values).
- A log-linear model for estimating the parameters of a high-order classification (contingency) tables.
- The model parameters of a high-order classification table. This is more or less Hammerlag's approach, as a difference with the former approach, the log-likelihood function is directly maximized without the classical logarithmic transformation.
- If covariates (continuous data) are going to be considered directly in the model, then the equations for parameter estimation of a generalized linear model have to be adapted to multi-Poisson distribution.

Time period T to be considered is a month or a year. From a mathematical point of view, calibration data are:

1. Dependent variable: number of incidents per road section in a time period T for a group of time periods, i.e. if T is a month then a 6 or 12 months should be considered for calibration purposes.
2. Independent variables: for each road section and time period T, a class for each factor involved in the model should be computed. The historic database must contain directly the needed data or a related type of data which allows the classification or aggregation for continuous data.
3. Reliability of data is not a problem in long term models since historic data stored in the database are supposed to be verified after being processed for the data collection module. Missing data in any of the variable would cause the elimination of the observation.

Parameters in the incident prediction model are estimated on the basis of the maximum likelihood method. Hammerslag proposal may be transformed to the generalized linear models frame.

Let $\boldsymbol{\lambda}_i = \boldsymbol{\alpha}_J \ell_i$ be the expected number of incidents per road section *j* of type *J* in a time period of length T, that is the Poisson parameter of the number of incidents distribution. Being α_J , the product of parameter estimates for factor values defining the J class.

Since the Poisson distribution is a type of distribution belonging to the exponential family (as the normal distribution), a generalized linear model with an expected number of incidents per road section type in a time period of length T being the product of factor parameter per road section length may be easily deduced, using as a link function the logarithmic function, the log-likelihood function must be maximized.

Let index *J* indicate the type of section, assuming the redefinition of a unique factor that incorporates all roadway, traffic and weather factors.

Let index *k* indicate the time period observation.

Let Y_i indicate an observation of a road section *j* in a time period *k* that belongs to type J.

Let $\log \boldsymbol{\lambda}_i = \boldsymbol{\eta}_i = X_i^T \boldsymbol{\beta}$ be the link function value for observation *i* expressed in vectorial form.

The iterative method of scoring process (on *m*) for the computation of the singular point of the log-likelihood function is defined as follows:

$$\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(\mathbf{m})}\mathbf{X}\mathbf{b}^{(\mathbf{m}+\mathbf{1})} = \mathbf{X}^{\mathbf{T}}\mathbf{W}^{(\mathbf{m})}\mathbf{z}^{(\mathbf{m})} \quad \textit{where} \quad \begin{array}{l} \mathbf{W}^{(\mathbf{m})} = \left(w_{ii}\right) \quad \textit{and} \quad w_{ii} = e^{\mathbf{X}_{\mathbf{i}}^{\mathbf{T}}\mathbf{b}^{(\mathbf{m})}} \\ z_i^{(m)} = \mathbf{X}_{\mathbf{i}}^{\mathbf{T}}\mathbf{b}^{(\mathbf{m})} + \left(y_i - e^{\mathbf{X}_{\mathbf{i}}^{\mathbf{T}}\mathbf{b}^{(\mathbf{m})}}\right) e^{-\mathbf{X}_{\mathbf{i}}^{\mathbf{T}}\mathbf{b}^{(\mathbf{m})}} \end{array}$$

Anyway, Hammerslag proposal for computing $\boldsymbol{\lambda}_i = \boldsymbol{\alpha}_J \ell_i$ estimates does not rely on generalized linear models and without considering the possibility of defining a link function that relate expected value for incidents in a section to a linear model of a set of parameters $\boldsymbol{\eta}_i = X_i^T \boldsymbol{\beta} = \ln(\boldsymbol{\lambda}_i) = \ln(\boldsymbol{\alpha}_J \ \ell_i)$ optimizes the log-likelihood function on $\boldsymbol{\alpha}_J$. This approach, by inspection of the resulting equations is quite similar to the multiproportional methods for estimating contingency tables, that appear frequently in O/D estimation matrices in transportation demand analysis.

Let us start assuming a matricial problem of dimension R^{nxn}. The cells in this matricial problem, arising frequently in the estimation of contingency tables, are for us the expected number of incidents in a given period of time T, for sections in the cross-class category defined by *i*-category of rows and *j*-category of columns. The number of categories of factor rows defines row dimension and the number of categories of factor columns defines column dimension.

Let us assume that the marginal row and column totals of the O/D matrix are known (marginal totals, the total number of incidents for each category of a factor). Let O_i denote the total number of incidents (row sum) of the category *i* of row factor, and Dj denote the total number of incidents (column sum) for the category *j* of column factor. We assume that a basic matrix that has to be updated according to the marginal totals and the basic matrix cells $\left\{d_{ij}\right\}$ consists of Poisson distributed integers. The method of maximum likelihood is applied then to estimate the updated matrix $\left\{D_{ij}\right\}$. Without any loss of generality, we can assume that $O_i > 0$ and $Dj > 0$ and that in order to assure feasibility, the sums of the marginal totals by row factor equals the marginal totals by column factor. The mathematical. programming problem may be formulated as,

$$\begin{array}{c} Max_{D_{ij}} \ \sum_i \sum_j d_{ij} \log D_{ij} \\ s.t. \\ \sum_j D_{ij} \quad = \quad O_i \quad \forall i \\ \sum_i D_{ij} \quad = \quad D_j \quad \forall i \\ D_{ij} \quad \geq \quad 0 \quad \forall i, j \end{array}$$

Let α^1 bee the dual variables associated with row factor constraints and α^2, those associated with the column factor constraints. Formulating the Lagrangean dual of the problem, the Kuhn-Tucker optimality condition for the problem yield to the solution values, for optimal dual variables:

$$\left\{ D_{ij} = {d_{ij}} \Big/ {\alpha_i^1 + \alpha_j^2} \right\}$$

This model resulting from the maximum likelihood approach in the case of Poisson distributed values is different from the maximum entropy model, also known as biproporcional balancing, Fratar or Furness method, which assume a multinomial distribution of cell values.

An algorithm that can be adapted to the current matrix estimation problem is:

Step 0:

Set α^1, α^2 to 0.

Step 1:

For each row category i, find $\overline{\alpha}_i^1 \geq -\min_j \alpha_j^2$ that minimizes

$$\left| \sum_j {d_{ij}} \Big/ {\overline{\alpha}_i^1 + \alpha_j^2} - O_i \right|$$

Set $u_i \leftarrow \alpha_i^1 - \overline{\alpha}_i^1 \quad and \quad \alpha_i^1 \leftarrow \overline{\alpha}_i^1$.

Step 2:

For each column category *j*, find $\overline{\alpha}^2_{ji} \geq -\min_i \alpha^1_i$ that minimizes

$$\left|\sum_i \frac{d_{ij}}{\alpha^1_i + \overline{\alpha}^2_j} - D_j\right|$$

Set $w_j \leftarrow \alpha^2_j - \overline{\alpha}^2_j \quad \textit{and} \quad \alpha^2_j \leftarrow \overline{\alpha}^2_j$.

Step 3:

If $\|u\| > \varepsilon \;\; or \;\; \|w\| > \varepsilon$ go to Step1, otherwise compute the optimal values

$$\left\{D_{ij} = \frac{d_{ij}}{\alpha^1_i + \alpha^2_j}\right\}$$

End Algorithm

The one-dimension minimization problems in steps 1 and 2 are equivalent to finding the zero of the corresponding function, if it exists; if the function has no zero-crossing in the considered interval, the minimum is always attained at the lower bound. The algorithm is shown to converge (6) and a proposal for solving efficiently the one-dimensional subproblems arising in steps 1 and 2 by Newton's method (successive linear approximation), results in the following recursion formulas in Step 1:

$$\left\{\overline{\alpha}^1_i = \overline{\alpha}^1_i + \frac{\sum_j \frac{d_{ij}}{\overline{\alpha}^1_i + \alpha^2_j} - O_i}{\sum_j \frac{d_{ij}}{\left(\overline{\alpha}^1_i + \alpha^2_j\right)^2}}\right\}$$

and in Step 2:

$$\left\{\overline{\alpha}^2_{ji} = \overline{\alpha}^2_j + \frac{\sum_i \frac{d_{ij}}{\alpha^1_i + \overline{\alpha}^2_j} - D_j}{\sum_i \frac{d_{ij}}{\left(\alpha^1_i + \overline{\alpha}^2_j\right)^2}}\right\}$$

The former algorithm can be extended to multiple classification by K factors (K>2). We denote by $M_{i_k}^k$ the marginal total for category i_k of factor *k*, which is assumed to be input data to the process.

General Algorithm for K classification factors (proposal, not theoretically proved):

Step 0:

Set $\boldsymbol{\alpha}^1,\ldots,\boldsymbol{\alpha}^K$ to **0**. Initialize $D_{i_1\ldots i_K}\leftarrow d_{i_1\ldots i_K}$.

Step 1:

For each category i_k of factor *k*

Find $\overline{\alpha}_{i_k}^k$ that minimizes

$$\left|\sum_{i_1}\cdots\sum_{i_{k-1}}\sum_{i_{k+1}}\cdots\sum_{i_K}\frac{d_{i_1\ldots i_K}}{\overline{\alpha}_i^k+\sum_{j\neq k}\alpha_{i_j}^j}-M_{i_k}^k\right|$$

Set $u_i^k\leftarrow\alpha_i^k-\overline{\alpha}_i^k \quad and \quad \alpha_i^k\leftarrow\overline{\alpha}_i^k$.

Step 2:

If $\left\|u^k\right\|>\varepsilon \quad for \quad any \quad k$ go to Step1, otherwise compute the optimal values

$$\left\{D_{i_1\ldots i_K}={d_{i_1\ldots i_K}}\Big/{\sum_j\alpha_{i_j}^j}\right\}$$

End Algorithm

The initialization step (Step 0) can be adapted to:

- $D_{i_1\ldots i_K}\leftarrow d_{i_1\ldots i_K}$, with initializes the expected values to a former estimates or
- $D_{i_1\ldots i_K}\leftarrow l_{i_1\ldots i_K}$, with initialized the expected values according to the weight of the cell, that is in our case, the lenght of the road section type or the product of covariates (quantitative data), as lenght, flow, etc.

In this former general approach, the expected number of incidents for each cross category are estimated. It is not necessary to define a design matrix X, as in the statistical approach of a generalized linear model, that is going to be of *huge dimensionality*: number of factors per summatory of the number of categories per factor minus the number of factors (assuming no interactions between variables). This approach avoids numerical problems arising in the iterative process of system resolution involving a huge *dense* matrix.

Prediction is enormously simplified with this former approach, since the interpretation of the parameter estimates gets easier since $\boldsymbol{\lambda}_i = \boldsymbol{\alpha}_J \ell_i$ is $D_{i_1 \ldots i_K}$ in the matrix estimation development.

4.4 Long-Term Estimation Model: Incidents Multinomial distributed.

The logit model developed for dichotomous response data appearing in a model for short term incident prediction is adapted in this section to polychotomous response variables, as it is the case in long-term estimation models where the number of incidents is modelled as a binomial distribution. Let us suppose that the response variable Y may take any of *m* qualitative values, which for convenience, are numbered *0* to *m-1*. Although the categories of Y are numbered, it is not necessary to attribute ordinal properties to these numbers (as it is the case when they are binomial distributed). Let $\Pi_{ij} = \mathrm{P}(Y_i = j)$ represent the probability that the *i*th observation falls in the *j*th response variable category.

A linear relationship of Π *to* X^T, a set of p regressors, is given by the symmetric form of the *multivariate logistic distribution function*:

$$\Pi_{ij} = \frac{e^{X_i^T \beta_j}}{\sum_{j=0}^{m-1} e^{X_i^T \beta_j}}$$

because $\sum_{j=0}^{m-1} \Pi_{ij} = 1$, it is necessary to impose a linear constraint on $\boldsymbol{\beta}_j$ to define them uniquely, $\sum_{j=0}^{m-1} \boldsymbol{\beta}_j = 0$. The log odds for any pair of categories *k* and *l* is a linear function of the difference between their parameter vectors:

$$\log\left(\Pi_{ik} \big/ \Pi_{il} \right) = X_i^T \left(\boldsymbol{\beta}_k - \boldsymbol{\beta}_l \right)$$

To fit model to data, the maximum likelihood method is invoked again. First, let us remark that each Y_i takes its possible values 0,1,...,m-1 with probabilities $\Pi_{i0},\dots,\Pi_{i(m-1)}$ and define dummy variables $W_{i0},\dots,W_{i(m-1)}$ so that $W_{ij} = 1$ if $Y_i = j$

and 0 otherwise.

Then the log-likelihood function is given by:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N}\sum_{j=0}^{m-1} W_{ij}\ X_i^T \boldsymbol{\beta}_j\ - \sum_{i=1}^{N} \log\left(\sum_{j=0}^{m-1} e^{X_i^T \boldsymbol{\beta}_j}\right)$$

Differenciating the log likelihood function with respect to the parameters and setting the partial derivatives to zero, produces the nonlinear estimating equations

$$\sum_{i=1}^{N} W_{ij}\ X_i\ =\ \sum_{i=1}^{N}\left(\frac{e^{X_i^T \boldsymbol{\beta}_j}}{\sum_{l=0}^{m-1} e^{X_i^T \boldsymbol{\beta}_l}}\right) X_i \qquad for\ \ j = 0,\dots,m-1$$

$$s.t.\ \ \sum_{j=0}^{m-1} \boldsymbol{\beta}_j = \mathbf{0}$$

The resulting vectors of parameters $\boldsymbol{\beta}_0,\dots,\boldsymbol{\beta}_{m-1}$ share the usual properties of maximum likelihood estimators.

The fitted probabilities $\hat{Y}_{ij}$ are given by the following equation:

$$\hat{Y}_{ij} = \frac{e^{X_i^T \boldsymbol{\beta}_j}}{\sum_{l=0}^{m-1} e^{X_i^T \boldsymbol{\beta}_l}}$$

Log linear models are models for the association among variables in a contingency table, when all variables are categorical, their joint sample distribution defines a cross classification or contingency table, where, in general, each combination of variable categories is observed more than once. Since most applications treat one variable as the dependent variable, log-linear models are generally applied only as a convenient means of fitting an equivalent logit model when all independent variables are qualitative (or categorical).

Let us treat in some detail contingency tables for two qualitative variables with *r* and *c* categories, respectively, which define a *r* x *c* contingency table. Let d_{ij} the observed frequency on cell *(i, j).* The marginal frequency on row *i* is notated d_{i+} and the marginal frequency on column *j* by d_{+j}, being n the total number of observations in the sample.

If it is assumed that the observations are produced by choosing an independent random sample of size *N* from a population characterized by probability Π_{ij} of selecting an observation in cell *(i,j),* then the expected frequency value of cell *(i,j)* is given by $D_{ij} = \mathrm{E}\left[d_{ij}\right] = N\,\Pi_{ij}$. It is generally simpler to estimate these expected frequencies than directly determine maximum-likelihood estimates of the parameters:

$$\log D_{ij} = \mu + \alpha_i^1 + \alpha_j^2$$

(Observe that cell estimates are related in a multiplicative way to an exponentian transformation of model parameters, as in Hammerslag proposal)

The likelihood function for estimating cell frequencies not model parameters and assuming a multinomial distribution of cell observations has a simple form:

$$L(\mathbf{D}) = -\sum_{i=1}^{r}\sum_{j=1}^{c}\left(D_{ij}\log\left(\frac{D_{ij}}{d_{ij}}\right) - D_{ij}\right) + 1$$

and it is called the maximum entropy model.

If we want to update a base matrix, observed frequencies, to satisfy marginal totals for rows and columns, the problem in the cell frequencies estimation form has to be restricted to a double set of linear constraints, totals by rows and totals by columns

$$\begin{aligned} &Min_{D_{ij}} \quad \sum_{i=1}^{r}\sum_{j=1}^{c}\left(D_{ij}\log\left(\frac{D_{ij}}{d_{ij}}\right) - D_{ij}\right) + 1 \\ &\qquad\qquad \sum_{i=1}^{r} D_{ij} = D_{+j} \quad \forall\, j \\ &.s.t. \quad \sum_{j=1}^{c} D_{ij} = D_{i+} \quad \forall\, i \\ &\qquad\qquad D_{ij} \geq 0 \quad \forall\, i, j \end{aligned}$$

There is a classical algorithm known as biproportional balancing, Fratar or Furness which determine an optimal solution:

$$D_{ij} = d_{ij}\, e^{\alpha_i^1 + \alpha_j^2} \qquad \alpha_i^1, \alpha_j^2 \;\; \textit{Lagrange multipliers}$$

The positive aspect of the former formulation it is the possibility of solving the problem by the well-known Kruithof method:

Kruithof Algorithm:

Input:

d_{ij}, D_{i+}, D_{+j}

Output:

$D_{ij} \;\; satisfying \; m\arg inal \; totals$

STEP 0: Initialization

Set $t \leftarrow 0$, $D_{ij}^t \leftarrow d_{ij}$, $r_i^t \leftarrow 1$, $s_j^t \leftarrow 1$.

STEP 1:

 for each row *i*

$$r_i^{t+1} \leftarrow r_i^t \frac{D_{i+}}{\sum_{j=1}^{c} D_{ij}^t} \;\; ; \quad D_{ij}^{t+\frac{1}{2}} \leftarrow D_{ij}^t \frac{D_{i+}}{\sum_{j=1}^{c} D_{ij}^t}$$

endfor;

for each column *j*

$$s_j^{t+1} \leftarrow s_j^t \frac{D_{+j}}{\sum_{i=1}^{r} D_{ij}^{t+½}} \;\; ; \;\; D_{ij}^{t+1} \leftarrow D_{ij}^{t+½} \frac{D_{+j}}{\sum_{i=1}^{r} D_{ij}^{t+½}}$$

endfor;

$t \leftarrow t+1$

STEP 2: Convergence Test

If Convergence Test is satisfied **STOP** otherwise **GOTO** STEP 1

End Algorithm

The method can be easily generalized to an iterative proportional fitting algorithm that provides maximum-likelihood estimates of the expected frequencies. Beginning with a multitable of ones or a previous table or just a weighted multitable representing the product of quantitative variables for each cell; the method successively adjusts the estimated expected frequencies to agree with each marginal table fit under a model. Adjustment for one such marginal generally disturbs agreement with the others. This procedure is repeated, however, until the estimated expected frequencies agree simultaneously with all marginals to be fit. Convergence takes place when the estimates stabilize to some preset level of precision, from one cycle of adjustments to the next.

K-Multiproportional Algorithm (not theoretically proved):

Input:

$$d_{i_1 \dots i_K}, \; D^*{}_{j+_k} = \sum_{i_1} \dots \sum_{i_{k-1}} \sum_{i_{k+1}} \dots \sum_{i_K} D^*_{i_1 \dots i_{k-1}\, j\, i_{k+1} \dots i_K} \; (m\arg inal\; total\; for\, category\; j\, of\;\; factor\, k$$

Output:

$$D^*_{i_1 \dots i_K} \;\; satisfying\;\; m\arg inal\;\; totals$$

```
STEP 0: Initialization
Set $t \leftarrow 0$, $D^{t}_{i_1 \ldots i_K} \leftarrow d_{i_1 \ldots i_K}$, $r^{k;t}_{j} \leftarrow 1 \ \forall k = 1 \ldots K$.

STEP 1:
 for each dimension $k$
     for each category $j$
          $r^{k;\,t+1}_{j} \leftarrow r^{k;t}_{j} \dfrac{D^{*}_{j+_k}}{D^{t+\frac{(k-1)}{K}}_{j+_k}}$ ; $D^{t+{(k)}/{K}}_{i_1 \ldots i_{k-1} j i_{k+1} \ldots i_K} \leftarrow D^{t+{(k-1)}/{K}}_{i_1 \ldots i_{k-1} j i_{k+1} \ldots i_K} \dfrac{D^{*}_{j+_k}}{D^{t+{(k-1)}/{K}}_{j+_k}}$
     endfor;
endfor;
 $t \leftarrow t+1$

STEP 2: Convergence Test
If Convergence Test is satisfied STOP otherwise GOTO STEP 1
End Algorithm
```

References

1. K.J.Kihlberg and J.K.Tharp; Cornell Aeronautical Lab. *Accident Rated as Related to Design Elements of Rural Highways*. NCHRP, Rept. 47 pp 173, 1968.
2. R. Hamerslag, J.P. Roos and M. Kwakernaak. *Analysis of Accidents in Traffic Situations by means of Multiproportional Weighted Poisson Model*. Transportation Research Record 847. 1982.
3. J.Ortúzar and L.G.Willumsen. Modelling Transport.John Wiley and Sons (1990).
4. INRESPONSE. Deliverable 4.1. Description of Task 4.3 Incident Prediction.
5. J.Fox. Linear Statistical Models and Related Methods with Applications to Social Research. John Wiley and Sons, 1984.
6. H. Spiess, *A maximum Likelihood model for estimating origin-destination matrices*. Transportation Research B, Vol 21B no.5, pp. 395-412, 1987.

5. IN-RESPONSE : BASIC PREDICTION MODEL

The basic IN-RESPONSE Incident Prediction model relates the expected number of accidents to road and traffic characteristics. The model has a multiplicative form and ***the expected number of accidents on a road section in a given period of time*** (the dependent variable) is a function of the estimated parameters and the length of the section studied.

The independent variables are all divided into classes; the number of factors and of their classes determines the number of coefficients to be estimated. The model equation is the following:

$$\mu_{klm} \;=\; E(Y_{klm,L}) \;=\; \beta_{1k} \cdot \beta_{2l} \cdot \beta_{3m} \cdot L_{klm}$$

where:

μ_{klm} = expected number of incidents on a section with characteristics $ß_{1k}$, $ß_{2l}$ and $ß_{3m}$ and length L;

L_{klm} = total length of the road section that belongs to the categories k,l and m;

$ß_{1k}...ß_{3m}$ = the coefficients (representing the different infuential factors $ß_1$, $ß_2$ and $ß_3$);

k..m = classes (each factor is divided into several classes; no continuous variables).

The probability of y_{klmn} incidents is:

$$P(y_{klm}) = \frac{e^{-\mu_{klm}} \mu^{y_{klm}}}{y_{klm}!}$$

To estimate the coefficients a, b, c and d, the log-likelihood function L^* is maximized:

$$L^* = \sum\sum\sum \ln P(y_{klm})$$

The maximal value of the log-likelihood can be found by setting the partial derivatives to zero:

$$\frac{\partial L^*}{\partial \hat{\beta}_{1k}} = \sum_l \sum_m (-\beta_{2l}\beta_{3m}...L_{klm}) + \sum_l \sum_m (\frac{y_{klm}}{\hat{\beta}_{1k}}) = 0, \forall k$$

and also

$$\frac{\partial \mathrm{L}^*}{\partial \hat{\beta}_{2\mathrm{l}}} = 0, \forall \mathrm{l}$$

$$\frac{\partial \mathrm{L}^*}{\partial \hat{\beta}_{3\mathrm{m}}} = 0, \forall \mathrm{m}$$

The coefficients b$_1$, b$_2$ and b$_3$ can be determined by solving the following set of (non-)linear equations:

$$\hat{\beta}_{1k} = \frac{y_{k...}}{\sum_k \sum_l \sum_m (\hat{\beta}_{2l}\hat{\beta}_{3m}\ldots L_{klm})}, \forall k$$

$$\hat{\beta}_{2\mathrm{l}} = \frac{\mathrm{y}_{.\mathrm{m}.}}{\sum_{\mathrm{k}} \sum_{\mathrm{l}} \sum_{\mathrm{m}} (\hat{\beta}_{1\mathrm{k}}\hat{\beta}_{3\mathrm{m}}\ldots \mathrm{L}_{\mathrm{klm}})}, \forall \mathrm{l}$$

$$\hat{\beta}_{3\mathrm{m}} = \frac{\mathrm{y}_{..\mathrm{m}}}{\sum_{\mathrm{k}} \sum_{\mathrm{l}} \sum_{\mathrm{m}} (\hat{\beta}_{1\mathrm{k}}\hat{\beta}_{2\mathrm{l}}\ldots \mathrm{L}_{\mathrm{klm}})}, \forall \mathrm{m}$$

with

$$\sum_{\mathrm{l}} \sum_{\mathrm{m}} \mathrm{y}_{\mathrm{klm}} = \mathrm{y}_{\mathrm{k}..}, \forall \mathrm{k}$$

$$\sum_{\mathrm{k}} \sum_{\mathrm{m}} \mathrm{y}_{\mathrm{klm}} = \mathrm{y}_{.\mathrm{l}.}, \forall \mathrm{l}$$

$$\sum_{\mathrm{k}} \sum_{\mathrm{l}} \mathrm{y}_{\mathrm{klm}} = \mathrm{y}_{..\mathrm{m}}, \forall \mathrm{m}$$

In the following section, the generic theory leading to the development of an iterative scheme for the resolution of the former equation set is presented. But another approach, much more intuitive is proposed for solving the estimation problem, it relies on the techniques for estimating contingency tables expected values, and it is widely employed in travel demand estimation (O/D matrices). First of all,let us consider a number of factors K affecting the number of expected incidents in a given section:

$$\mu_{i_1 \ldots i_K} = \beta_{i_1} \cdots \beta_{i_K} l_{i_1 \ldots i_K} \quad \forall\, i_1 \ldots i_K$$

where $l_{i_1 \dots i_K}$ is the length of the section (group of sections) defined by the cross category $\left(i_1 \dots i_K\right)$ or a previous estimate of the expected number of incident for cells, or the product of continuous variables related to cells.

Let $\mu_{j+_k} = \sum_{i_1} \cdots \sum_{i_{k-1}} \sum_{i_{k+1}} \cdots \sum_{i_K} \mu^*_{i_1 \dots i_{k-1}\, j\, i_{k+1} \dots i_K}$ notate the total number of incidents for category *j* of factor *k*, this is the marginal total of factor *k*, for $k = 1, \dots, K$. *Marginal totals are assumed to be an input to* the process, and *the output are* the right hand side terms in the definition equation $\mu_{i_1 \dots i_K} \quad \forall i_1 \dots i_K$,; additionally, vector parameters $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K$ are also computed.

Let $\hat{\mu}^{(t)}_{i_1 \dots i_K}$ define the estimated number of incidents for a given period T in a section in the cross category $\left(i_1 \dots i_K\right)$ at iterate *t*; and $\hat{\mu}^{(t)}_{j+_k} = \sum_{i_1} \cdots \sum_{i_{k-1}} \sum_{i_{k+1}} \cdots \sum_{i_K} \hat{\mu}^{(t)}_{i_1 \dots i_{k-1}\, j\, i_{k+1} \dots i_K}$ the estimated marginal total for category *j* of factor *k* at the *t-th* iteration.

Let N be the total number of incidents occurred during the time interval under study.

The proposed method successively adjusts the estimated expected frequencies to agree with each marginal table fit under a model. Adjustment for one such marginal generally disturbs agreement with the others. This procedure is repeated, however, until the estimated expected frequencies agree simultaneously with all marginals to be fit. Convergence takes place when the estimates stabilize to some preset level of precision, from one cycle of adjustments to the next.

K-Multiproportional Algorithm:

Input:

$l_{i_1 \dots i_K}, \ \forall \left(i_1 \dots i_K\right) \ and \quad \mu_{j+_k} \ (m\mathrm{arg}inal\, total \ for\, category \ j\, of \ \ factor\, k)$

Output:

$\hat{\mu}^*_{i_1 \dots i_K} \ \forall \left(i_1 \dots i_K\right) \ satisfying \ m\mathrm{arg}inal \ totals$

STEP 0: Initialization

Set $t \leftarrow 0$, $\mu^{(t)}_{i_1 \dots i_K} \leftarrow l_{i_1 \dots i_K}$, $r_j^{k;t} \leftarrow 1 \ \ \forall k = 1 \dots K$.

STEP 1:

for each dimension *k*

```
        for each category $j$

            $r_j^{k;\,t+1} \leftarrow r_j^{k;t} \frac{\mu^*_{j+_k}}{\hat{\mu}_{j+_k}^{\left(t+\frac{(k-1)}{K}\right)}} \;;\quad \hat{\mu}_{i_1\ldots i_{k-1} j i_{k+1}\ldots i_K}^{\left(t+{(k)}/{K}\right)} \leftarrow \hat{\mu}_{i_1\ldots i_{k-1} j i_{k+1}\ldots i_K}^{\left(t+{(k-1)}/{K}\right)} \frac{\mu^*_{j+_k}}{\hat{\mu}_{j+_k}^{\left(t+{(k-1)}/{K}\right)}} \quad \forall i_1\cdots i_{k-1}, i_{k+1}\cdots i_K$

        endfor;

    endfor;

    $t \leftarrow t+1$

    STEP 2: Convergence Test

    Let $\beta_{i_k}^{(t)} \leftarrow r_i^{k;t} \;\; \forall\left(i_1\ldots i_K\right) \forall\, k=1,\ldots,K$
    If $\left\|\vec{\beta}_k^{(t)} - \vec{\beta}_k^{(t-1)}\right\| \le \varepsilon$ is satisfied for $k=1,\ldots,K$    STOP
                                        otherwise GOTO STEP 1

    End Algorithm
```

The algorithm can be coded in any programming language (C, Fortran) and it is simple to understand how it proceeds.

Programming and parameter calibration have to be performed before the end of IN-RESPONSE project by December 97.