Statistical modelling and uncertainty analysis of data on accidents, flow and emission

Urban Hjorth
Department of Mathematical Statistics, Chalmers,
41296 Göteborg, Sweden

February 18, 2003

Vinnova project 1156-4022, DNR 2001-03881; (KFB 1999-0260) Final report.

Abstract: Three different aspects of vehicle traffic, that have been modelled and analysed statistically, are reviewed in this project overview. The extraction of subflow proportions and travelling time distributions from traffic counts, the estimation of relative collision safety of different vehicles from data on injury classes in collisions, and environmental questions related to measures of vehicle emissions and catalyst function. Methods for estimation and uncertainty evaluation from complex data are at focus.

Key Words: Travelling time, dynamic OD estimation, car crash, driver safety, vehicle emission, catalyst aging, certification, safety margin, light beam measurement, parameter estimation, bootstrap, McMC, likelihood.

1 Background

This project started with grants from the former Swedish Transport & Communications Research Board (KFB) for the period 1999 – Feb. 2003, and was the third in a series of projects on statistics in transport research with an open formulation of its goals. In the reorganisation of research boards, the project was inherited by Vinnova in 2001. This report summarises research since 1999. We are grateful to Vinnova and KFB for their support.

In the work we have cooperated with The Swedish Road and Traffic Research

Institute (VTI), Volvo Truck, IVL, the Department of Chemistry and the Environmental Section at Chalmers and have also used data from SCB.

2 Information on traffic flow from counts

Although it is obvious that traffic counts give direct information about flow at the measurement spot, the challenge in our approach has been to get more details about subflows between different measurement places without costly extra equipment. Our basis of information is the variability itself, in other words, the randomness of the traffic flow. Part of the fluctuations at one measurement spot will persist at downstream measurement spots and produces dependency between the counts. This brings information about the subflow between the measurement places in terms of both the proportions using that link and the distribution of travelling times. This opens new possibilities for a more measurement based estimation of traffic streams in a network where not every link has to be measured or individual vehicles followed. This can be a complement and sometimes a substitute to origin-destination (OD) estimates based on optimisation with entropy maximisation, often made without direct empirical input from flows.

Our modelling approach belongs to a tradition of dynamic OD estimation dating back to the eighties, with Cremer and Keller (1987) as one early reference. It has been tried on well-defined objects like for example crossings. An introduction to this tradition is given in Hjorth (1999) and will not be repeated here. Let us only mention that a modelling rather close to ours was made by Jarret and Wright (1990) but seems to have overlooked an important filtering when applied to real data.

2.1 Modelling the counts

Discretise time with a time step that can be chosen, depending on distances and traffic volume, somewhere between a second up to a few minutes. In our case the original data were in continuous time so any time step was available. Under the assumptions of 1^o independent driver behaviour and 2^o uncongested network, the relation between traffic counts $X(t)$ at one place A and $Y(t)$ at another (downstream) place B, is written

$$X(t) = \sum_{\tau=0}^{m} U_\tau(t) + X_1(t), \tag{1}$$

and

$$Y(t) = \sum_{\tau=0}^{m} U_\tau(t-\tau) + Y_1(t). \tag{2}$$

Here $U_\tau(t)$ is the subflow passing A at time t and B at $t+\tau$, $X_1(t)$ denotes the flow at A which will not pass B and $Y_1(t)$ is the flow at B having other origins than past A. We also assume that Y_1 is independent of X and U. In this setting we derive the covariances in Hjorth (1999) as

$$Cov(X(t), Y(t+\tau)) = \sum_{\tau'} p(\tau') Cov(X(t), X(t+\tau-\tau')), \tag{3}$$

where $p(\tau)$ is the probability that a driver at A will be counted at B after τ time units. When counts at A are uncorrelated in time, the sum simplifies to a single term which gives

$$p(\tau) = \frac{Cov(X(t), Y(t+\tau))}{Var(X(t))}, \tag{4}$$

otherwise the preceding equation will give the system of equations from which a (finite) vector of $p(\tau)$ can be solved from estimated variances and covariances for the flow. They can also be solved from a corresponding regression equation. The estimation part uses a certain time period and in principle the situation should be stationary for the estimates to represent more momentaneous covariances but this is not fulfilled for real traffic. Instead systematic fluctuations in the total volume of traffic flow may cause "false" covariation in the estimates if not properly met. This is a pitfall where earlier efforts have failed. Our solution is to use local estimation of flow levels, regard these as time varying expected values and estimate covariances from the remaining fluctuations. We can regard this as local kernel regression estimate or as a filter taking out low frequency variation and keeping the informative high frequency part. Details of this filtering are in Hjorth (1999, 2002) where also the estimates are illustrated. In Hjorth (2002) the methods above are generalised to more general stochastic processes and applied to a more complex traffic link with traffic lights and crossings. The distribution of travelling times and route choice was estimated by the corresponding probabilities $p_{ij}(\tau)$ for flow between positions i and j. Since traffic lights were not synchronised with our time steps, their phase varied during the estimation period and the traffic lights did therefore not destroy the estimates of the proportions and the travelling time distribution (where we include waiting in the travelling time). In this formulation the modelling has similarities with signal processing models using filters on the data and filter descriptions of the traffic elements.

2.2 Modelling classified counts

Our discussion so far has assumptions about uncongested flow and independent driver behaviour, which means that the vehicles should have very little interaction. At higher traffic intensities, other factors like bottlenecks at crossings will determine the traffic fluctuations and mask the covariation due to vehicles driving between the measurement points. In Bergendorff (1999) the subflow problem was analysed based on a classification of the vehicles. Any such classification will do in principle, the one available was in terms of vehicle length measured as distance between first and last axes. Instead of counting just the numbers, she showed how the fluctuating proportions of vehicle types could be utilised in subflow estimation. The basic inference idea was to condition on flow at A, total flow at B and derive either a conditional likelihood or the moments for classified flow at B. Some different estimation methods were compared and two of them, one weighted moment estimator and one conditional maximum likelihood estimator could be recommended.

2.3 Uncertainty analysis of subflow estimates

Knowing the uncertainty of complex parameter estimates is an important part of the inference. Usually the uncertainty estimation is harder than the parameter estimation itself. Analytical uncertainty estimates are possible under a Poisson process modelling of the traffic flow. This was utilised in Hjorth (1999) to derive a covariance matrix for the estimated $p(\tau)$-vector, $\tau = 0, \ldots, m$..

$$C_{\hat{p}} = AC_{\eta}A', \tag{5}$$

where

$$A = \begin{bmatrix} \alpha(1) & \alpha(2) & & \alpha(n) & 0 & 0 & 0 \\ 0 & \alpha(1) & & \alpha(n-1) & \alpha(n) & 0 & 0 \\ \vdots & & & & & & \\ 0 & 0 & \alpha(1) & & & & \alpha(n) \end{bmatrix}, \tag{6}$$

$\tilde{X}(t)$ is $X(t)$ minus the estimated expected value or in other words the output from the high pass filter, $\alpha(t) = \tilde{X}(t)/\sum_1^n \tilde{X}(s)^2$,

$$C_{\eta} = \sum_{t=1-m}^{n+m} X(t)(C)_t + diag(Var\,(Y_1(\cdot))) \tag{7}$$

where finally $(C)_t$ is a (large) matrix having the matrix $C = -pp' + diag(p)$ in position (t, t) and upwards and zeros in the other positions.

Asymptotic variance/covariance results give simpler expressions which are useful for large amounts of data and as guidelines in the planning of data volumes needed for successful estimates. Under conditions stated in Hjorth (1999) the variance converges as

$$n\,Var(\hat{p}(i)) \to \frac{p(i) - p(i)^2}{\lambda_x} + \sum(p(s) - p(s)^2) + \frac{\lambda_{Y_1}}{\lambda_x}. \tag{8}$$

Here λ_x denotes the (mean) traffic intensity of the flow x. Based on the variance expressions, confidence intervals for the estimated route selection/travelling time parameters was also derived.

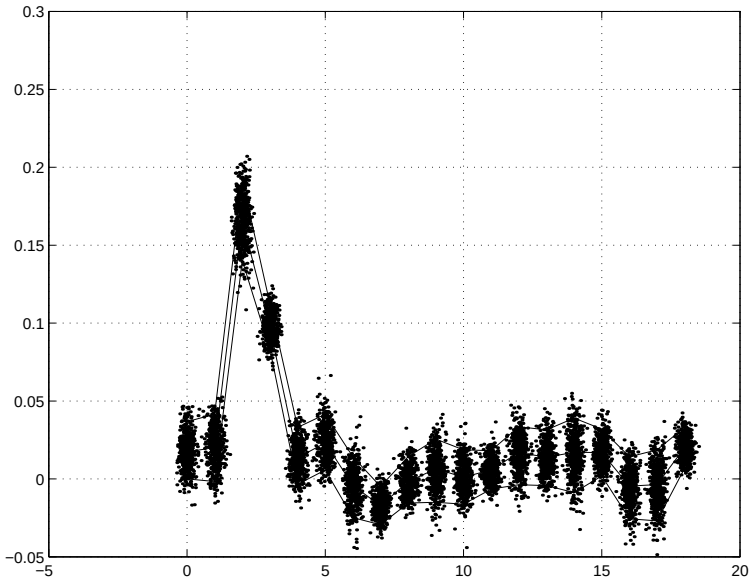

Figure 1: Bootstrapped estimates of travelling probabilities as function of travelling time.

For the more general modelling, analytical variance/covariance analysis seems out of reach. Instead computer intensive methodology comes in handy and bootstrap (Efron-Tibshirani 1993, Hjorth 1994) is the most convenient tool. This means that we resample data from the original data in ways that will keep the important random variation and the dependency of the original data. In Bergendorff (1999) dependency in time was handled by a block bootstrap where the original data were seen as blocks of a certain length and the resampling draw at random the starting times of such blocks until as much data were collected as in the original set. Repeated estimates on these data give the

bootstrap estimated variability of the estimates. In Hjorth (2002), data from several days were collected and were seen as independent from day to day. Instead of the block resampling, now a high level of resampling was possible using the entire dayly data as resampling units. This will keep all dependencies within the same day of the data and the bootstrap variability estimates the variability of the parameter estimates and produces confidence intervals. For an example, see Fig. 1.

3 Modelling and analysing accident data

Traffic safety problems can be approached from at least two points of view. One is to reduce the amount of accidents, in particular the most serious ones, and the other is to limit the consequences of accidents that occur. Vehicles' inner safety and their protection of driver and passengers is an important part of the second approach. Information about this safety is available from true accidents, but since speed and circumstances are usually very badly known, the information is typically hard to extract. Another source of information is collision testing in controlled experiments. However, there is no one to one relation between such results and the statistics from true accidents with a much larger variation of speed, hitting points and forces and with humans involved. It is therefore of high interest to extract as much information as possible from the accidents that do occur and to explore statistical tools for doing this.

3.1 Vadeby's collision model

In data on real collisions, the forces are usually unknown but we know at least that both colliding vehicles are exposed to the same forces and can therefore be compared statistically. This requires a modelling with relevant parametrisation of the vehicle safety and the forces and since every collision brings in new parameters, the statistical analysis has a theoretical challenge in terms of a vector of so called incidental parameters which is growing in proportion to the number of data. In the thesis work, Vadeby (2003), different types of analyses with incidental parameters and estimation of the parameters are studied together with methods for uncertainty evaluation. In the application to accident data, all these analyses are based on the same collision model given in Vadeby (1998). This model uses a basic random process, a so called birth model, to describe the probability that a driver ends up in different injury classes. This process starts in the state "unhurt", and moves at random into more serious damage. Depending on the collision forces and the safety param-

eter a driver will spend a certain "time" in this process. Since the collision forces are unknown, we only have information about the ratio of these "times" for the two colliding drivers. The approach in Vadeby (1998) is to define incidental parameters representing the forces or equivalently the times spent in the damage process. If this parameter is given for one of the drivers, the other driver is given a value adjusted for the ratio of vehicle weights and the ratio of the vehicles' safety parameters. We use here that heavy vehicles experience less change of speed than light vehicles in a collision between them, according to the law $m_1 \triangle v_1 = m_2 \triangle v_2$ where m is mass and $\triangle v$ is change of speed. With this model, an explicit expression for the likelihood can be given. Since the model has three types of parameters: the safety parameters, the parameters of the injury class process, and the vector of collision forces, where each parameter type can be estimated relatively easy by numerical methods if the other two are known, it follows that iteration of the three numerical steps will give the maximum likelihood estimates.

In Vadeby (2000) the model is extended with information about driver's age and sex, since these factors are known to affect the risk of death and probably also the probability for the other injury classes in the same direction. If some vehicle types had very different driver populations in these respects, the relative safety parameters could be misleading without parameters allowing for this. Now a fourth type of parameters is included to take care of the driver characteristics and iterations are over four steps instead of three. Only minor effects on the estimated safety parameters was observed in this set of data, but the possibility to allow for such factors is demonstrated.

By this modelling, estimates and confidence intervals for the relative safety are given for a number of frequent car makes, see some examples below. In the interpretation of results we must remember that the safety parameter has been adjusted for vehicle weight and the comparison therefore represents equal change of speed (retardation) for the vehicles during collisions.

3.2 Estimation uncertainty

Since the likelihood estimation is an iterative solution, it has no closed and explicit form. However, the likelihood itself can be expressed and differentiated. Due to the growing vector of incidental parameters, standard asymptotic tools, based on likelihood derivatives and the information inequality (see Lehmann-Casella, 1998) are not useful for the variance of our maximum likelihood estimates. One possible approach to this problem is bootstrap analysis where the entire estimation is repeated on data resampled from the original data. For the estimation of parameters from data on injury classes, a certain

number of collisions is needed with each considered car model. Data involving very rare models were therefore excluded early on. In a resampling of collisions, the problem may happen again in the resampled data and the measures of variability could then represent the variations in number on top of the variations in accident outcome. This was avoided by resampling conditional on approximately the same number of collisions as in the original data for each car model. In practise, collisions with less frequent cars were resampled first in order to keep their numbers and the most frequent car make (in Sweden Volvo) was allowed to vary somewhat in order to compensate for this. The entire bootstrap analysis is quite computer intensive but works and gives estimated variances and confidence intervals demonstrating that some car makes are significantly better than others in their protection of the driver. Some examples are

Comparison	90 % bootstrap interval
Volvo rel. Vw	$0.19 < \alpha < 0.58$
Saab rel. Vw	$0.24 < \alpha < 0.84$
Audi rel. Vw	$0.18 < \alpha < 0.91$.

Here $\alpha < 1$ means safer than Vw given the same change of speed. In a later analysis (Vadeby 2003b), using a subset of the data with only collisions between five vehicle makes, the first two estimates changed to

Volvo rel. Vw	$0.19 < \alpha < 0.72$
Saab rel. Vw	$0.17 < \alpha < 0.97$

showing that also collisions with other makes will contribute to the precision of the estimates.

3.2.1 Alternatives for deterministic incidental parameters

In Vadeby (2002a, 2003a) the discussion is extended to two other methods based on the profile likelihood and the delta method, After an overview of the rather rich literature she compares uncertainty estimates for these methods and the bootstrap. Since the collision data are tricky to handle, these methods are investigated in a simpler Poisson model of similar type i.e. with incidental parameters and indirect integer valued data coming in pairs. Here the different approaches can be theoretically analysed and compared. In this model (Y_{i1}, Y_{i2}) are independent $Po(\theta_i)$ and $Po(\alpha\theta_i)$ given the incidental parameter θ_i. Estimates of α are studied for both deterministic and random θ_i and without conditioning on any sufficient statistics for θ_i in order to keep the similarity with car crash models. Putting in estimates for θ and neglecting their uncertainty in an information matrix will vastly underestimate the

uncertainties about α. Using instead the profile likelihood

$$L_p(\alpha) = \max_{\theta} L(\alpha, \theta)$$

and the asymptotic variance relation

$$Var(\hat{\alpha}_p) = -(E[\frac{\partial^2 \log L_p}{\partial \alpha^2}])^{-1}$$

we get the result $(\alpha^2 + \alpha)/\sum \theta_i$ (or rather that the denominator times the variance converges to $(\alpha^2 + \alpha)$. Vadeby then shows that the same result is achieved asymptotically by the bootstrap in spite of an induced dependency there and also that the delta method, based on series expansion and the central limit theorem, gives convergence to a normal distribution with the same asymptotic variance in spite of the complication that the estimate is a ratio without mean and variance for finite sample. The same properties up to first order are seen for random incidental parameters as for deterministic.

The similarity between variance estimates for the different approaches gives support for an assumption that the same equivalence could be valid also in the collision model. The analysis can be seen as a theoretical support for the bootstrap results already discussed as well as for the other methods.

3.3 Bayesian models and McMC

The growing vector of incidental parameters, which represents the unknown forces in the collisions, can be avoided by a different modelling where these values are seen as independent random variables from some parametrised distribution. In principle this is nice since the parameters are now a fixed vector of reasonable dimension. However, this does not give attractive computations in the Vadeby model as long as the other parameters are still deterministic. A computationally better analysis is achieved in a full Bayesian model. This approach is made on both the Poisson model and on the larger collision model. Now prior distributions must be defined as probability densities for all the parameters. How this is made using Gamma distributions, inverse Gamma distributions and hierarchical (two level) models is described in Vadeby (2003b). The structure is designed as to not impose conclusions that are not indicated by the data themselves. These prior densities multiplied with the likelihood, i.e. with the probability of our data as function of the parameters, gives the posteriori probabilities of the parameters up to a proportionality constant (which is impossible to integrate out). This is where another computational method, using Markov chain simulations, is the ideal tool for our needs.

3.3.1 Markov chain monte carlo

When we have a multivariate probability density known up to a normalising constant, it is possible and surprisingly simple to create a Markov chain having this density as its stationary distribution and jumping around ergodically so that the time spent in different regions will after long time be proportional to the stationary probability of the region. The Bayesian posteriori distribution is conveniently studied in this way. See e.g. Gilks et al. (1996). In practise, only low-dimensional marginal distributions will converge fast enough and usually one- or two-dimensional parameter distributions are searched but the simulations may take place in high dimension and many low-dimensional distributions can be studied from the same series of simulations.

Since the Markov chain can be defined in many different ways, especially for a high-dimensional problem, some chains are more effective than others because they converge faster. In order to optimise the chains we need measures of convergence. The more advanced measures need analytical knowledge about the target distribution in order to give bounds for the speed of convergence. Often such knowledge is not at hand and the more general criteria must be based on the simulation output itself. Reviews of such criteria are given in Brooks and Roberts (1998), Cowles and Carlin (1996) and Mengersen et al (1999). Since the existing convergence indicators were either too informal for optimisation of the simulations or had other disadvantages for our application in their definitions or performance, we decided to develop a new measure in Hjorth-Vadeby (2002). This is based on the well-known Kullback-Leibler distance between distributions and was designed to compare the distribution of subsequences in the simulation with the result for the entire simulation or to compare parallel simulations with the joint result. This measure seems to work very well and has a series expansion where the leading term can be interpreted as the relative uncertainty (σ/μ) for cell frequencies in the subsequences. When the convergence is acceptable, this leading term also dominates the measure in the cases we have studied and this makes the measure easy to interpret. The new KL-measure compares favourably to Gelman and Rubin's (1992) method with parallel chains and its generalisation by Brooks and Gelman (1998) and also to Yu and Mykland (1998) who base their method on the behaviour of cumulative sums. In Vadeby (2003b) this measure is used to optimise the proposal distributions used in the definition of her Markov chains (by Metropolis Hastings method). The measure is also used as the prime diagnostic for the convergence of the McMC analyses based on both Gibbs and Metropolis-Hastings methods. From the analyses we get simulated posteriori-distributions for the parameters of interest and in particular, credible intervals for the most probable values of the relative safety parameters.

4 Models related to vehicle emission.

Under this heading we will report one work on analyses of catalysts, another on environmental certification problems in vehicle engine production and a third work on remote measurements of vehicle emissions. All three in cooperation with other departments or companies.

4.1 Catalyst aging analysis

The chemistry around catalysts is studied in detail by several research groups. However, the functioning of used cars' catalysts is a statistical mixture of different types of catalysts and different uses of the cars and is therefore not a simple function of basic chemical knowledge. For that reason, data on used catalysts were collected from scrapped cars of different ages and makes. This collection was made by the Department of Inorganic Environmental Chemistry. Our contribution was modelling and synthesis of these data in terms of light off temperatures, i.e. the rather distinct temperatures where the catalysts start to work efficiently, and the steady state properties at typical working conditions. Test equipment and background is described more fully in Chan (2000) and to some extent also in Hjorth et al. (2002), where otherwise the modelling and handling of these data is at focus.

For light off temperatures of CO, HC, and NO, a modelling with regression on age was written as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{1}$$

where y is light off temperature, x is age (minus average age) defined as driving distance in 100000 km,

$\beta_0 = \beta_0(m,u) = \beta_{00}(m) + \beta_{01}(m)u,$

$\beta_1 = \beta_1(m,u,v) = \beta_{10}(m) + \beta_{11}(m)u + \beta_{12}(m)v,$

$\beta_2 = \beta_2(m),$

and m means car make, u the model year, v the kilometres driver per year.

The model was tested against sub-models with simpler structure and some of the parameters were excluded before most of the analyses were made. For details about models and results see the reference above. Since, for practical

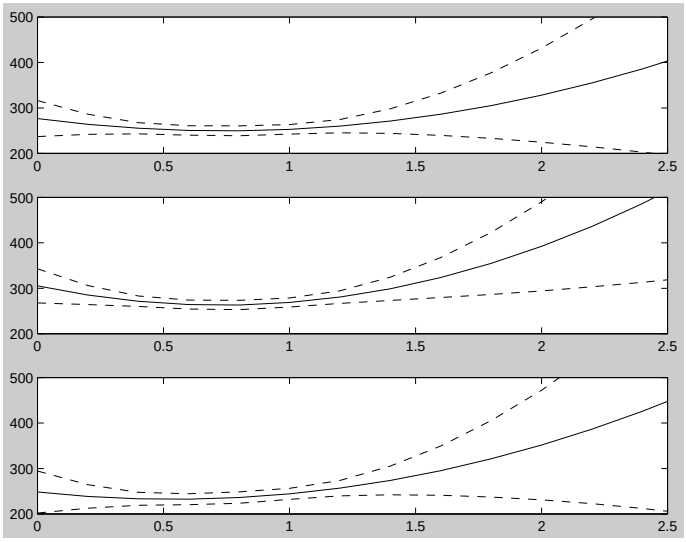

Figure 2: Light off temperature estimates and 95% conf. int. as function of driving distance x in 100000 km. Results for the average catalyst in data. CO, HC, NO from above. From Environmetrics 13, 2002, p. 840.

reasons, it was impossible to sample catalysts from a well-defined population, a new summary concept was defined as "the average catalyst in data". For this concept the aging function is illustrated with confidence limits in the following figure from Environmetrics (Hjorth et.al. 2002).

In a multivariate regression based on the full model above, the dependency of light off temperatures for CO, HC, NO, was illustrated by an estimated correlation matrix:

$$\hat{\rho} = \begin{pmatrix} 1 & 0.60 & 0.95 \\ 0.60 & 1 & 0.53 \\ 0.95 & 0.53 & 1 \end{pmatrix}.$$

In particular CO and NO had a high and positive correlation in this analysis where the deviations (residuals) from the regression equation are the correlated objects. Notice that this correlation measures if catalysts are simultaneous good or bad for the different substances emitted, and do not describe variation within the same catalyst, where correlations of different sign may occur.

For steady state measurements, the character of data motivated a modelling in terms of categories. Using car makes and driving distance as predictors, a generalised regression model for the category probabilities was defined. Significant degradation with age, in terms of increased probability for worse category, was shown for CO and NO but not for HC. The overall impression of this in-

vestigation was, however, that used catalysts keep their performance better than expected up to the ages of around 150000 km driving distance. If higher vehicle emissions with age are observed, during this period, other parts of the engine regulation system may contribute to this. For longer driving distances the collected data were not very informative.

4.2 Environmental certification

Certification of (diesel) engines on different markets is a challenge for vehicle producers in particular since the emission limits are tightened over time. The engine producer, Volvo, takes basic measurements on every produced unit including a few environmentally related data like fuel consumption and soot. More detailed environmental data are, however, complicated and resource consuming and are therefore only taken on a very limited sample of engines from the production. Based on these results, the company must convince authorities about the emissions quality and must also be prepared to prove for example by a sequential test procedure, that the requirements are met. Since the burden of proving this is on the producer, it is necessary to have a certain margin to the limits. With a too small such margin, the sequential test may take many costly testing units and in worst case fail to be accepted.

The statistical challenge is now to make inference from the small samples of environmental data to the entire population of thousands of produced engines and decide which margin is needed in the samples in order to have an appropriate margin for the whole production with high probability. This analysis can also give information about the sampling of test units in terms of how they should be selected and their number. Analyses of the available data from the entire production showed that data from different vehicles were not independent but had a time series dependency which was not very strong but still of some importance for the problem. Together with Volvo Truck, a project was defined on these problems and some more specific ones not discussed here. The project ended up in the report Nilsgren (2002).

4.2.1 Learning from other variables

The important environmental variables were only available from a relatively small number of sampled units. The sampling had not been random but in groups of a few almost simultaneously produced engines taken quarterly. The good thing with this was that variation within groups and variation between groups could be compared and give information about the time dependency

of the data. The most critical substance in this study was NO_x and part of the modelling concentrated on this substance.. It turned out that one of the variables that were measured on every unit had very similar estimated relations between neighbours and distant (in time) engines and did also look similar in its distribution. This was a power variable which had also good correlation to the NO_x if low power was associated with high emission. Instead of a theoretical modelling (which was first at mind) it seemed like a more close empirical model was available in this variable measured on thousands of units. Here the dependency and natural variation of the production process was automatically included. The problem was therefore analysed as if these data were the environmental values. Different sampling plans were compared by simulation in this set of data. The effect of different margins based on samples from the long series of data could also be simulated on these data. The margins were of course based on measures from the sampled units only in order to imitate the uncertainty of a real situation.

This study gave recommendations to sample smaller groups of engines well spread out in time. Statistically optimal for the precision of a margin was to take one unit at a time, but pairs of engines was only slightly less efficient and had advantages in terms of continued information on dependency. The benefit of sampling enough engines was also clearly demonstrated.

4.3 Analysis of remote vehicle emission data

Light beam technique can be used to measure the exhausts from vehicles in real traffic. The physical basis of this method is the absorption of light at certain substance-specific wavelengths in a beam crossing the traffic lane and mirrored back. The measurement technique is named FEAT (fuel efficiency automobile test) from another type of application and has been introduced in Sweden for remote vehicle emissions by IVL who have made rather extensive measurements over the last ten years, see for example Sjödin-Lenner (1995). Due to the turbulence, the exhausts move very randomly and the amount crossed by the beam is not in itself a relevant measure. Instead all other absorptions are related to the absorption due to the release of carbon dioxide, and we get measured proportions of different substances. Since the air is not clean between the vehicles, the measurements have to be calibrated for the background levels, and only the increase will be registered. Calibration programs come with the instrument and have so far been accepted as they are. Hopefully they give unbiased values, however, there are several sources of random error, and one consequence of this is that some percent of the observations are negative.

In Bokelund (2003) such data together with observations of vehicle number plates, vehicle speed and acceleration have been analysed. Statistical models of different types have been put forward. From the number plates much details about the vehicles is available such as weight, age, catalysts etc. Regression models give estimated expected values of the exhaust proportions as functions of known variables. The uncertainty is also analysed and the significance of differences between vehicle groups and of speed and acceleration and weight and their combinations are studied by both regression and other technique.

For a subset of the measured vehicles we have more than one observation. This gives an opportunity to measure how representative one observation is as a predictor of later observations on the same vehicle. The result can be seen as a combined measure of both the stability in exhaust proportions over time for a vehicle and of the randomness in the measurements. For the deviations (residuals) from the estimated expected values it was seen that on average 73%, 46% and 31% of the CO, NO_x, and HC proportions did remain from one measurement to the next. An optimal prediction of future levels should therefore shrink the observations towards the regression estimated values by these factors.

The true emission levels (relative the CO_2 levels) have to be positive. They are also small and far below any theoretical upper limit so they can be modelled by a statistical distribution on the entire positive axis from 0 to infinity having its mass concentrated at low values. The log normal, the Weibull and the Gamma are examples of possible such distributions and the Beta or Pareto are useful alternatives if the upper limit is finite. In Bokelund (2003) the observations are modelled as Gamma distributed true levels observed with random and normally distributed measurement errors. The errors are postulated to have expected value zero and constant variance for all emission levels of the same substance. In the Gamma distribution, the expected values are supposed to follow the estimated regression function. In order to have a manageable parameter estimation, one of the two parameters in the Gamma distribution is held fixed (at an estimated value) and the other will be proportional to the expected value taken from the regression function. The estimated measurement uncertainty was rather large according to this modelling. However, since many assumptions are made, modelling errors may be part of the explanation for this result and further investigation of measurement error seems motivated and can be made both from a theoretical point of view on existing data and from an experimental point of view. The modelling used so far gives an interesting starting point for the theoretical approach.

5 Exchange of knowledge

Under this heading we will cover publications, seminars and conferences and the like. Our project has also influenced traffic research at VTI, at the Section of Traffic planning, Lund, and at the Department of Infrastructure, KTH, where analysis of uncertainty by computer intensive methods is now more widely used.

5.1 Conferences and seminars

Hjorth, U.: On modelling car catalyst degradation. ECAS, Garpenberg, Sept. 6-10, 1999.

Hjorth, U.: Estimating traffic sub-flows from filtered counts. 3rd KFB Research Conference, Stockholm, June 13-14, 2000.
Internet publication www.kfb.se/conf/trans2000.

Hjorth, U.: On space-time covariance for geostatistical data. Ties/SPRUCE 2000, Sheffield, Sept. 4-8, 2000.

Environment and Statistics Workshop, Stockholm, Nov. 28-29, 2000. Spatial problems, emission and spread were at centre in this meeting with main speakers from France and Finland and contributions from VTI, IVL, FFA and university departments. www.math.chalmers.se/Centres/SC/envstat2000. Chair of program committee and main organiser, U. Hjorth.

Transportforum, Linköping, Jan. 10-11, 2001. U Hjorth, A Vadeby.

Vadeby, A.: Vad säger kollisionsdata om säkerheten i fordon. (What is collision data telling us about safety in vehicles?) Transportforum, Linköping, Jan 11, 2001.

Hjorth, U.: A Markov model of speed and gap in rural traffic. Trafikdage på Aalborg Universitet, 27-28 Aug. 2001.

Hjorth, U.: Opponent to the PhD-thesis: A Bayesian approach to retrospective detection of change-points in road surface measurements by F. Thomas. Stockholm Sep. 24, 2001.

Hållbar utveckling i Sverige och globalt; Miljökonferens Göteborg Nov. 22, 2001. U Hjorth.

Vadeby A.: Inference in models with incidental parameters. Linköping University, Sept 19, 2001.

Umeå Statistical Winter Conference. (Topic: Mixed Effects Models), Mars 10-15, 2002. U. Hjorth.

Nordstat, Stockholm, June 9-13, 2002. U. Hjorth, A, Vadeby.

Vadeby, A.: Estimation with incidental parameters- comparing Monte Carlo and classical techniques by an example. Nordstat, Stockholm, June 11, 2002.

Vadeby, A.: Basic concepts, convergence diagnostics and applications of Markov chain Monte Carlo methods. Linköping University, Jan. 22, 2003.

5.2 Publications in the project

Bergendorff, P. (1999). Extracting information on traffic patterns from classified traffic counts. Thesis no 762, Linköping Studies in Science and Technology. (Lic.)

Bokelund, S. Statistisk analys av emissionsdata från vägfordon. (Statistical analysis of emission data from vehicles on the road. In Swedish with English summary.) Dept. of Mathematical Statistics, Chalmers—GU, 2003. (In cooperation with IVL.)

Hjorth, U. (1999). The inherent precision of regression estimated route probabilities. Transp. Res. B 33, 593–607.

Hjorth, U. (1999). Flow estimation and accident modelling. Studies in applied probability and statistics 1999:5, Dep. of Mathematical statistics, Chalmers/GU. Final report of the preceding project.

Hjorth, U. (1999). On space-time covariance for geostatistical data. Preprint 1999–64 Department of Mathematical Statistics, Chalmers/GU

Hjorth, U., (2002) Traffic subflow estimation and bootstrap analysis from filtered counts, Transportation Research Part B 36, pp. 345-359.

Hjorth, U., Chan, A., Zhao, D., Ljungström, E. (2002). Statistical analysis of catalyst data. Environmetrics 13, 831-846.

Nilsgren, M. Search of prediction models and sampling plans for emission tests on heavy diesel engines. Dept. of Mathematical Statistics, Chalmers—GU,

2002. (In cooperation with Volvo Truck.)

Vadeby, A. (2000) Including driver characteristics in a model of relative collision safety. LiTH-MAT-R-2000-19, Linköping University.

Vadeby, A. (2002b) Estimation in a Model with Incidental Parameters, LiTH-MAT-R-2002-02. Linköping University.

Vadeby, A. (2003a) Computer based statistical treatment in models with incidental parameters-inspired by car crash data. PhD thesis scheduled for May 9, 2003. Linköping University.

Included work:

Vadeby, A. (1998) Modelling and Inference of Relative Collision Safety in Cars, Anna Vadeby, LiU-TEK-LIC-1998:20. Linköping University.

Vadeby, A. (2000) Including Driver Characteristics in a Model of Relative Collision Safety, LiTH-MAT-R-2000-19. Linköping University.

Vadeby, A. (2002a)Estimation in a Model with Incidental Parameters, LiTH-MAT-R-2002-02.

Hjorth, U., Vadeby, A. (2002) The Empirical KL-measure of MCMC convergence, (submitted).

Vadeby, A. (2003b) On Gibbs Sampler and Metropolis-Hastings applied to Pairwise Poisson and Car Crash Data.

Vadeby, A. (2002b) Modelling of Relative Collision Safety Including Driver Characteristics, (submitted).

5.3 References to other work

Brooks, S.P, Gelman, A. (1998) "General Methods for Monitoring Convergence of Iterative Simulations" Journal of Computational and Graphical Statistics Vol 7, Nb 4, 434-455.

Brooks, S.P, Roberts, G.O (1998) "Assessing Convergence of Markov Chain Monte Carlo Algorithms" Statistics and Computing 8, 319-335.

Chan, A. (2000) Full scale testing of field-aged automotive catalytic converters. Dept. of Inorganic Environmental Chemistry, Chalmers, Göteborg.

Cowles, M.K., Carlin, B.P (1996) Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review, Journal of the American Statistical Association, Vol.91, No 434, 883-904.

Efron and Tibshirani (1993) An introduction to the bootstrap. Chapman and Hall/CRC.

Gelman, A, Rubin, D.B.(1992) "Inference from Iterative Simulation Using Multiple Sequences" Statistical Science, Vol 7, No 4, 457-511.

Gilks, W.R, Richardson, S, Spiegelhalter, D.J.(1996) "Markov Chain Monte Carlo in Practice", Chapman and Hall.

Hjorth, U. (1994) Computerintensive statistical methods — Validation, model selection, bootstrap. Chapman and Hall/CRC. Lehmann, E.L. (1983). Theory of Point Estimation, Wiley.

Lehmann, E.L., Casella, G. (1998). Theory of Point Estimation, Springer.

Mengersen, K.L, Robert C.P, Guihenneuc-Jouyaux, C (1999) " MCMC Convergence Diagnostics: A Review." Bayesian Statistics 6, J.M. Bernado, J.O. Berger, A.P. Davis and A.F.M. Smith (Eds.), pp 415-440. Oxford University Press, Oxford.

Sjödin, Å, Lenner, M. (1995). On-road measurements of single vehicle pollutant emissions, speed and acceleration for large fleets of vehicles in different traffic environment. Science of the total environment 169, 157.

Yu, B. Mykland, P. (1998) "Looking at Markov Samplers through Cusum Path Plots: a Simple Diagnostics Idea" Statistics and Computing 8, 275-286.