Time Series models for count data

Dimitris Karlis
Department of Statistics, Athens University of Economics, GREECE,
Email: karlis@aueb.gr

and

Tom Brijs
IMOB, University of Hasselt, BELGIUM,
E-mail: tom.brijs@uhasselt.be

Outline

- Intro and Motivation
- Some Models
- The Model of Zeger
- INAR Model
- Application
- Concluding Remarks – Further Topics

Caution: This presentations includes a number of equations!

Count Data

Count Data are encountered in:

- Accident analysis, epidemiology, sports, marketing, environmetrics, economy, finance, ecology, physics, biostatistics etc.
- Certain continuous data can be transformed to count data (e.g. rainfall data, threshold models etc)
- In many cases continuous data are truly measured as counts (e.g. time to unemployment)

Why time series models?

- In many circumstances the data are really time series (e.g daily or hourly or monthly number of crashes)
- Preliminary analysis has shown that in many datasets there is autocorrelation present, time series models can account for this
- Ignoring this autocorrelation the derived effects (coefficients of the model) may not be correct and thus the reported effect can be wrong (or at least we may report larger effects than those really exist)
- Note that autocorrelation can be in many cases due to the fact that our observations share the same conditions (e.g. environmental or weather conditions)

Why not standard time series models?

Standard time series models are based on an assumption of normal and related distributions which, while reasonable for continuous data, fail substantially for count data.

They fail for one or more of the following reasons:

- Low counts, small mean
- A lot of zeros
- Symmetry not present
- Probabilities hard to compute and interpret

Normal approximations exist and work well for large values (so, usually in aggregated data).

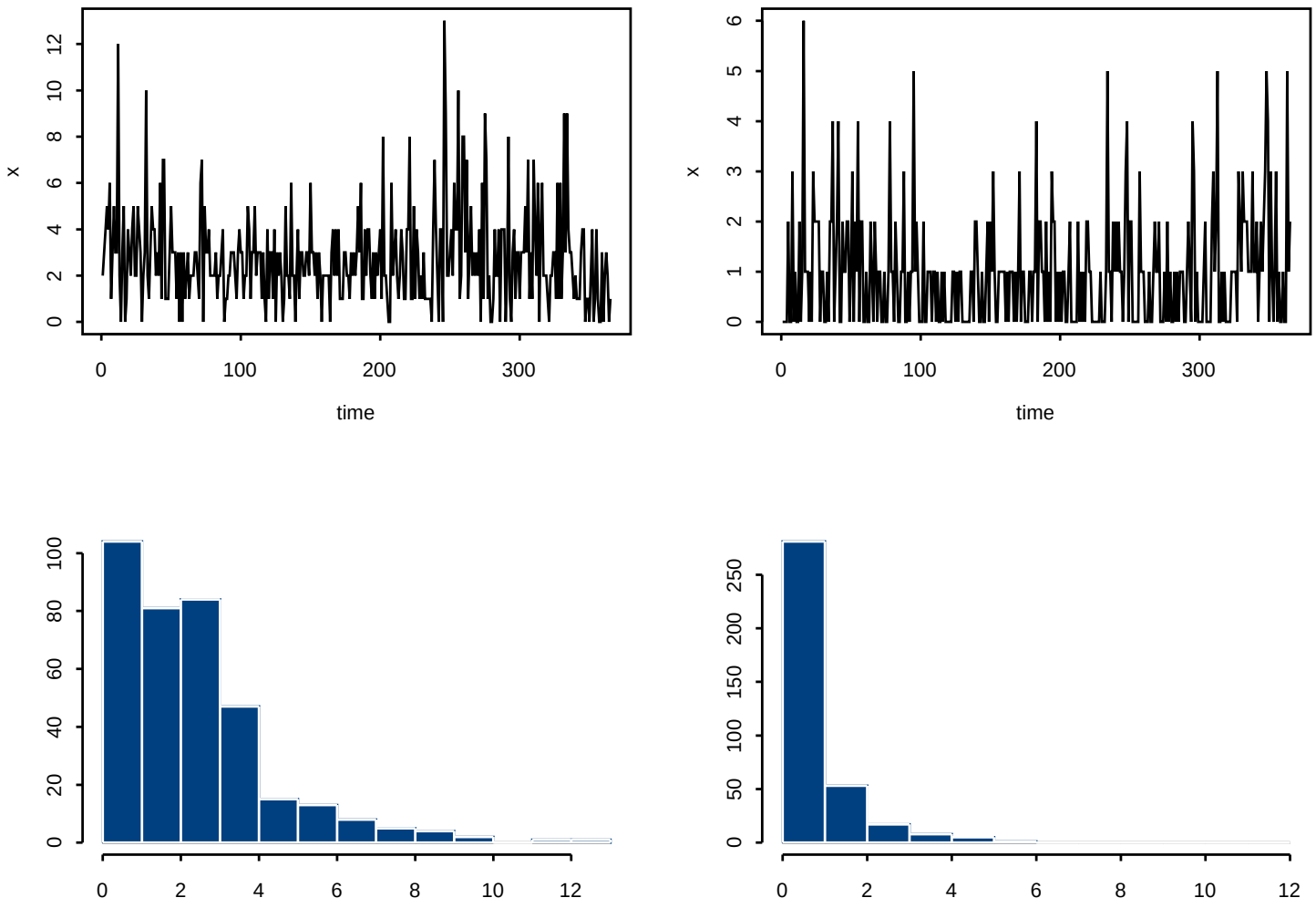

Figure 1: Example of discrete-valued time series

Time Series Models for Count Data

- Parameter driven

$$Y_t \mid \epsilon_t, x_t \sim Poisson(\epsilon_t \exp(x_t'\beta))$$

  ϵ_t is a latent process usually of a standard form as in classical time series, e.g $\varepsilon_t = \rho\varepsilon_{t-1} + w_t, \quad w_t \;\; \text{iid} \;\; N(0, \sigma^2)$

- Observation driven

$$Y_t \mid Y_{t-1}, x_t \sim Poisson(\exp(x_t'\beta + f(Y_{t-1})))$$

  where $f(y)$ can be any function

- Hidden Markov models : there is a hidden process that changes the state of the series and adds autocorrelation
- Integer Autoregressive Models: They mimic standard autoregressive models suitably defined for integers
- Other models

Integer Autoregressive

We mimic the classical AR model for normally distributed data. The process is defined as

$$Y_t = a \circ Y_{t-1} + R_t$$

where R_t is a sequence of uncorrelated non-negative integer-valued random variables having mean μ and finite variance σ^2 and X_0 represents an initial value of the process while the operator " $\circ$ " denotes the binomial thinning operator.

The operator " $\circ$ " is defined by

$$\alpha \circ X = \sum_{t=1}^{X} Y_t,$$

where Y_t are Bernoulli random variables with $P(Y_t = 1) = \alpha = 1 - P(Y_t = 0)$, $\alpha \in [0, 1]$ and is called the binomial thinning operator.

Properties

The mean and variance of a stationary INAR(1) process (i.e. $0 < \alpha < 1$) are constants given by the formulae

$$\mu_X = E(X_t) = \frac{\mu_R}{1-\alpha} \quad \text{and} \quad \sigma_X^2 = Var(X_t) = \frac{\alpha\mu_R + \sigma_R^2}{1-\alpha^2}, \tag{1}$$

where μ_R and σ_R^2 are respectively the (assumed finite) mean and variance of the i.i.d. innovations. The auto-covariance function of a stationary INAR(1) process $\{X_t\}_{t\in Z}$ is given by the formula

$$\gamma_X(k) = Cov(X_t, X_{t-k}) = \alpha^{|k|}\sigma_X^2 \quad , \quad k \in Z. \tag{2}$$

From the auto-covariance function, it is easy to obtain the autocorrelation function $\rho(k)$ as follows:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \alpha^{|k|}.$$

Thus, the autocorrelation function $\rho(k)$ decays exponentially with lag k.

Poisson Models

$$Y_t = a \circ Y_{t-1} + R_t$$

$$R_t \sim Poisson(\lambda_t)$$
$$\log \lambda_t = x_t' \beta$$

where x_t' are covariates associated with the t observation and β, as usual, the coefficients to be estimated. The autocorrelation in lag 1 is α.

A serious limitation is that marginally the data follow a Poisson distribution which is not so realistic. We may add covariates by assuming

$$\log\left(\frac{\alpha_t}{1-\alpha_t}\right) = \mathbf{w}_t'\gamma,$$

but this might cause problems in the interpretability of the model!

How to estimate the model?

Despite the complicated form of the model an easily programmable EM algorithm is available.

E-step: Using the current values of the estimates, say $\theta^{old} = (\beta^{old}, \gamma^{old})$, calculate

$$s_t = E(R_t \mid x_t, x_{t-1}, \theta^{old}) = \sum_{z=0}^{\infty} z \frac{P(R_t = z)P(Y_t = x_t - z)}{P(x_t|x_{t-1}, \alpha_t^{old}, \lambda_t^{old})}$$

for $t = 1, \ldots, T$, where $P(R_t = z)$ and $P(Y_t = x_t - z)$ are the probability functions of a Poisson and a binomial distribution respectively and $\lambda_t^{old} = \exp\left(\mathbf{z}_t \beta^{old}\right)$ and $\alpha_t^{old} = \exp(\mathbf{w}_t \gamma^{old})/(1 + \exp(\mathbf{w}_t \gamma^{old}))$.

The conditional expectation of Y_t given the data and the current values of the estimates can be determined by simple subtraction, as

$$c_t = E(Y_t|x_t, x_{t-1}, \theta^{old}) = x_t - s_t.$$

M-Step: Update the parameters in θ by fitting two GLM models. Namely, update β by fitting a Poisson regression model with response variables c_t and design matrix $\mathbf{z}$, while γ can be updated by fitting a binomial logit model with response s_t and design matrix $\mathbf{w}$.

Stop iterating when some convergence criterion is satisfied, otherwise, go back to the E-step.

Remark: The algorithm just described, ignoring the time series structure is in fact an algorithm that fits a Poisson-Binomial regression model.

Overdispersion

Poisson INAR model assumes that marginally the densities are Poisson which is too restrictive. We need to generalize so as to cover overdispersion. The situation becomes more complicated. It can be seen that the overdispersion of the series relates to the overdispersion of the innovation by

$$ID\left(Y_t\right) = \frac{a + ID\left(R_t\right)}{1 + a}$$

Hence, we need an overdispersed distribution for the innovations!

A simple idea is to use the negative binomial distribution but this leads to very complicated likelihood. Another idea is to use a finite mixture of Poisson distributions (Pavlopoulos and Karlis, 2006 hopefully). Estimation is feasible. There is a trade off between the descriptive and the predictive power of such a model.

Extensions

- Not constant α. Flexible but it allows for limited overdispersion.
- Define INAR(p) models of the form

$$Y_t = \sum_{i=1}^{p} a_i \circ Y_{t-i} + R_t$$

- Other distributional choices for R_t so as to provide tractable likelihoods!
- Dependence between Y_{t-1} and R_t.

Zeger's Model

We assume

$$\begin{aligned} Y_t &\sim Poisson(\mu_t \epsilon_t) \\ \log(\mu_t) &= x_t' \beta \\ \epsilon_t &= \phi \epsilon_{t-1} + R_t \\ R_t &\sim N(0, \sigma^2) \end{aligned}$$

Note that we asume that ϵ_t has an $AR(1)$ process. We may use any other process to replace it.

Pros: Overdispersion and correlation are imposed together by ϵ_t.

Cons: Hard to be estimated

Properties

$$\rho(k) = \frac{\rho_\epsilon(k)}{[(1+(\sigma^2\mu_t)^{-1})(1+(\sigma^2\mu_{t+k})^{-1})]^{1/2}}$$

where $\rho_\epsilon(k)$ is the autocorrelation of the process assumed for ϵ, hence, autocorrelation from the innovation process. Stationarity is not guaranteed.
Important things

- Good news: Since the correlation properties are coming from the innovation one can use any of the models for continuous time series to create a similar model for count data.
- Bad news: Estimation quite complicated: Use a GEE approach or an MCEM approach.

Hidden Markov Models

Consider that there are two states. The series can be in one of the states. Conditionally on these states the observations are independent. But since we cannot know the states the series is at each time point and because the states are correlated we observe data that are correlated because there is the latent process that changes the states at each time point!

Usually the states are connected through a Markov Process with some transition probabilities, we may allow for higher order Markovian properties etc.

Properties

Pros

- Easy interpretation
- Markovian and well understood properties

Cons

- Estimation is difficult
- The number of states is another thing that we have to take into account and estimate.
- Being Markovian we cannot have a large range of correlation structure

More Models

There are several other models proposed in the literature

- Models based on mixing - ARMA type models
- Dynamic Ordered Probit models for count data
- Composite models
- ...

Application - Summary

Aim: Examining effect of weather conditions on accident counts.

Data: Yearly crash count data on the major road network surrounding 3 major cities in the Netherlands. Weather data from nearby meteorological stations

Model: INAR model, with model selection technique to select variables.

Things to be considered: Explanatory weather variables are correlated. For each characteristic more than one variable available (causes multicollinearity). Transformations needed to avoid linear relationships. Day is used as a driver for different traffic volumes.

Limitations: Data are aggregated within days, so weather effects are somewhat smoothed out.
More details in Brijs, Karlis and Wets, (2006+)

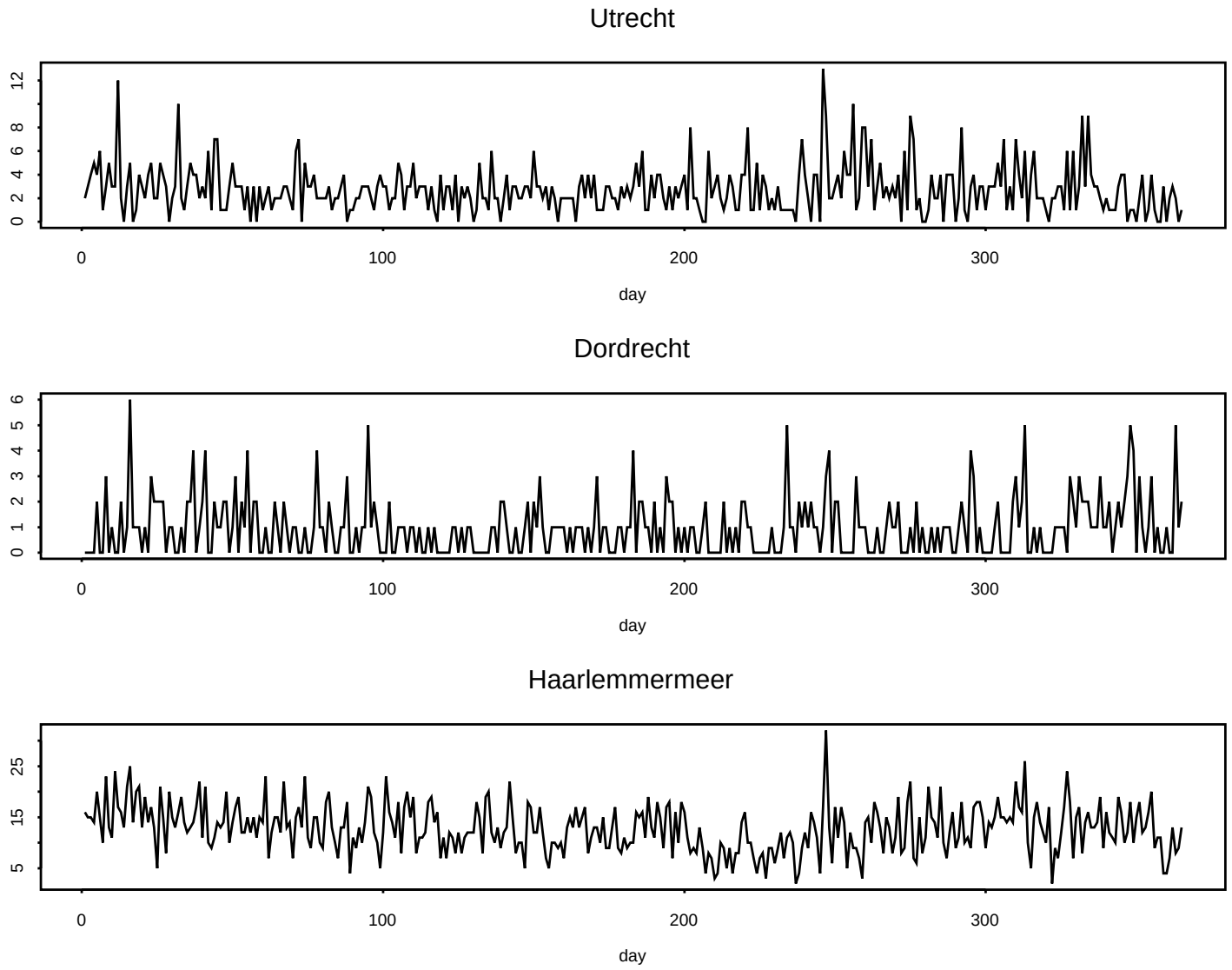

Figure 2: Accident counts for the three sites (daily data for 2001)

site	mean	variance	autocorrelation	variance/mean
Utrecht	2.747	4.227	0.0276	1.54
Dordrecht	0.950	1.239	0.0956	1.30
Haarlemmermeer	12.819	21.950	0.222	1.71

Table 1: Descriptive measures for the 3 series

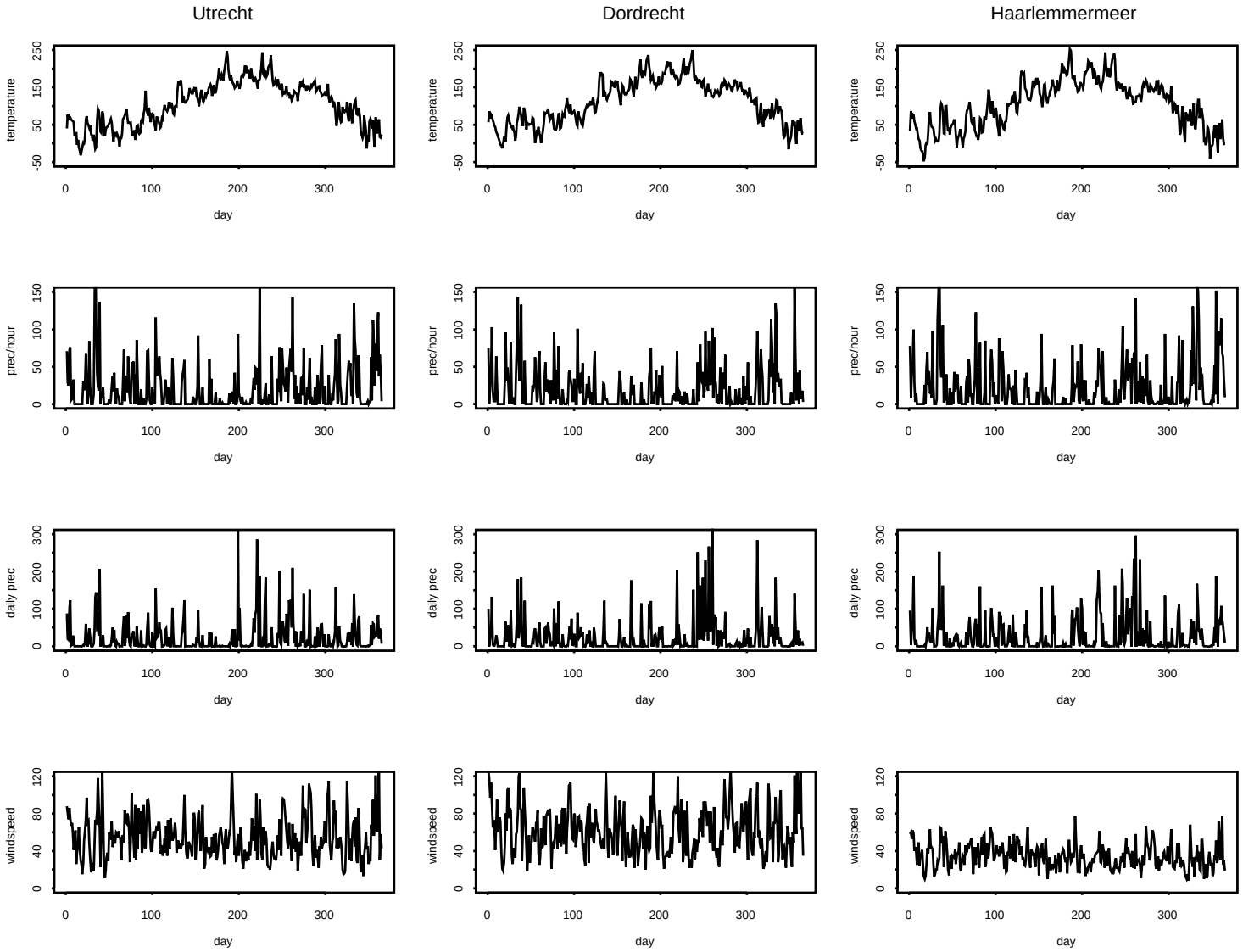

Figure 3: Plot of some of the weather variables for the three sites. There are notably different weather conditions

Results

	coefficient	Std. err	t-value	p-value
Constant	1.70479	0.10718	15.905	0.000
Utrecht	-1.38973	0.06835	-20.331	0.000
Dordrecht	-2.51865	0.08515	-29.578	0.000
Temperature			23.926	< 0.001
< 0	0.52376	0.11347	4.615	< 0.001
$[0, 10]$	0.321407	0.08256	3.892	< 0.001
$[10, 20]$	0.251626	0.07979	3.153	0.000
> 20	0	-	-	-
Dev of mean temp	0.00098	0.00057	1.703	0.081
Precipitation duration	0.00268	0.00050	5.344	0.000
Precipitation Intensity	0.00320	0.01430	2.240	0.025
Sun dazzle	0.19477	0.07903	2.464	0.013
% of max. possible sunshine duration	0.00116	0.00063	1.823	0.068

Continued

Day of the week			108.95	0.000
Monday	0.34479	0.05856	5.887	0.000
Tuesday	0.44668	0.05742	7.778	0.000
Wednesday	0.26593	0.05901	4.506	0.000
Thursday	0.28859	0.05866	4.919	0.000
Friday	0.43635	0.05698	7.657	0.000
Saturday	0.08122	0.06225	1.304	0.192
Sunday	0	-	-	-
Autocorrelation parameters				
Utrecht	0	0.03751	0	1
Dordrecht	0.07586	0.04317	1.757	0.078
Haarlemmermeer	0.14032	0.03912	3.586	0.003

Table 2: Results based on the fitted model INAR regression model

Model	Log-likelihood
Poisson regression	-2246.496
Negative Binomial Regression	-2243.034
Poisson INAR regression (INAR1)	-2242.629
Poisson INAR regression with covariates on α (INAR2)	-2238.961

Table 3: Comparison of different competing models

Ongoing Work

- Data from more routes and stations are now examined.
- Data cover different time spans (even hourly data).
- Weather data contain more info.
- More models are fitted and compared to verify the effects.
- Meta-analysis is used to combine the results.

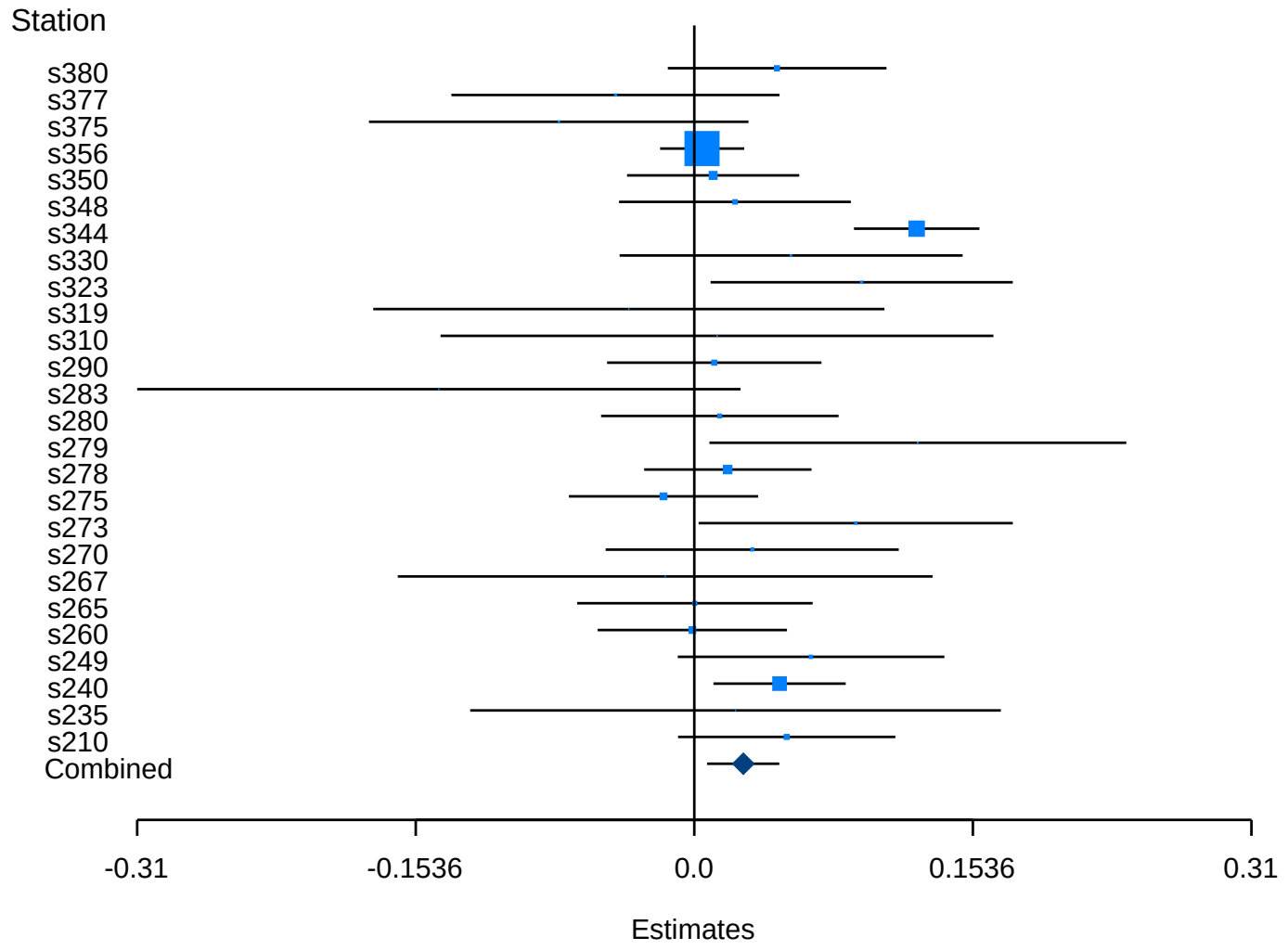

Figure 4: Intensity of precipitation

Concluding Remarks

- The literature of appropriate models increases.
- There is no need any more to use approximations based on continuous data.
- There are several models not treated here.
- Computational techniques can make the models easily available to the research community.

Work in Progress

- Overdispersed INAR models
- Bayesian estimation
- Define time series via the fahsionable copula models
- Compare different models: do they agree? or why do they disagree?

THE END

or just a beginning?