State estimation of Covid-19 disease in Denmark

Kasper Schou Telkamp

2022-12-23

Contents

Introduction

Early detection of outbreaks with communicable diseases are of great importance in order to initiate timely interventions and help prevent disease spread. When dealing with vast amount of data, automated procedures can supplement traditional surveillance methods, and help achieve earlier detection of the outbreak. Ultimately, leading to a reduction in the size of the disease outbreak.

In this report non-normal mixed effects models is evaluated on their ability to identify outbreaks of Covid-19 disease using data over new hospital admissions with Covid-19 in Denmark. Different implementations of generalized linear mixed models (GLMMs) in R packages is compared. Namely, the `glmmTMB` (Brooks et al., 2017) and `KFAS` (Helske, 2017) R packages available at Comprehensive R Archive Network (CRAN).

Data

In this project, the daily record of new hospital admissions with Covid-19 in Denmark grouped by region of residence and totals are used. This report is based on data from 1st of March, 2020 to 6th of November, 2022. The head and tail of the processed data are listed in Table 1.

Table 1: Processed data containing the daily record of new hospital admissions with Covid-19 in Denmark grouped by region of residence and with totals.

Dato	Hovedstaden	Sjælland	Syddanmark	Midtjylland	Nordjylland	Ukendt.Region	Total
2020-03-01	1	0	0	0	0	0	1
2020-03-02	0	0	0	0	0	0	0
2020-03-03	1	0	0	0	0	0	1
2020-03-04	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...
2022-11-03	20	12	9	10	2	1	54
2022-11-04	13	15	6	3	3	0	40
2022-11-05	7	6	6	3	3	0	25
2022-11-06	8	10	7	4	2	0	31

The data is publicly available and were obtained from Statens Serum Institut (SSI) website[1]. SSI collects the data from the National Patient Registry (NPR), which contains information about outpatient contacts from Danish public as well as private hospitals. The data from NPR has some delay. Therefore, the inventory is updated daily with real-time data from the regions. The regions provide snapshot-data twice to SSI daily at 7am and 3pm. A hospital admission related to Covid-19 is defined as an admission, where a patient is admitted within 14 days after a positive SARS-CoV-2 test. Patients that are tested positive for SARS-CoV-2 during an admission is also registered as a Covid-19 related admission. Furthermore, admissions with Covid-19 are only registered for patients that are present in at least one snapshot, or if the patient have been admitted for more than 12 hours according to NPR. The total number of new admissions to the hospital in Denmark are visualized in Figure 1.

[1] https://covid19.ssi.dk/overvagningsdata/download-fil-med-overvaagningdata

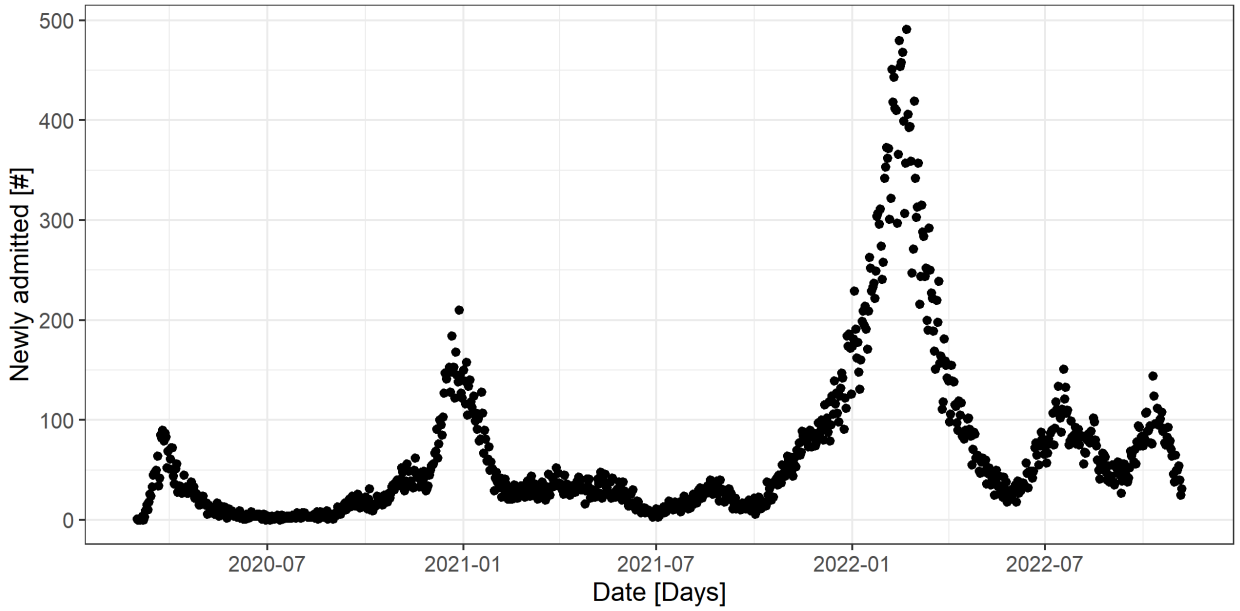

Figure 1: Total number of new hospital admissions in Denmark.

Clearly, the observed number of new hospital admissions are correlated in time. Consecutive epidemiological waves of Covid-19 disease have been observed, but with varying peak number of new hospital admissions. The peak number of new hospital admissions depends upon a number of influencing factors such as:

- Severity of illness
- Population in which the virus spreads
- How many are infected

Whereas the first two waves of new admissions to hospitals with Covid-19 disease can be explained by the severity of the illness and the population in which the virus spread, the large wave observed in early January, 2022, is more like likely caused by a vast amount of disease spread and the number of individuals that were infected at the same time.

Denmark is grouped into five regions: *Hovestaden*, *Midtjylland*, *Nordjylland*, *Sjælland*, and *Syddanmark*. If a patient is admitted to the hospital with Covid-19, but does not reside in any of the regions, they are marked as *Ukendt Region*. In Figure 2 the daily number of new hospital admissions grouped by region of residence is visualized.

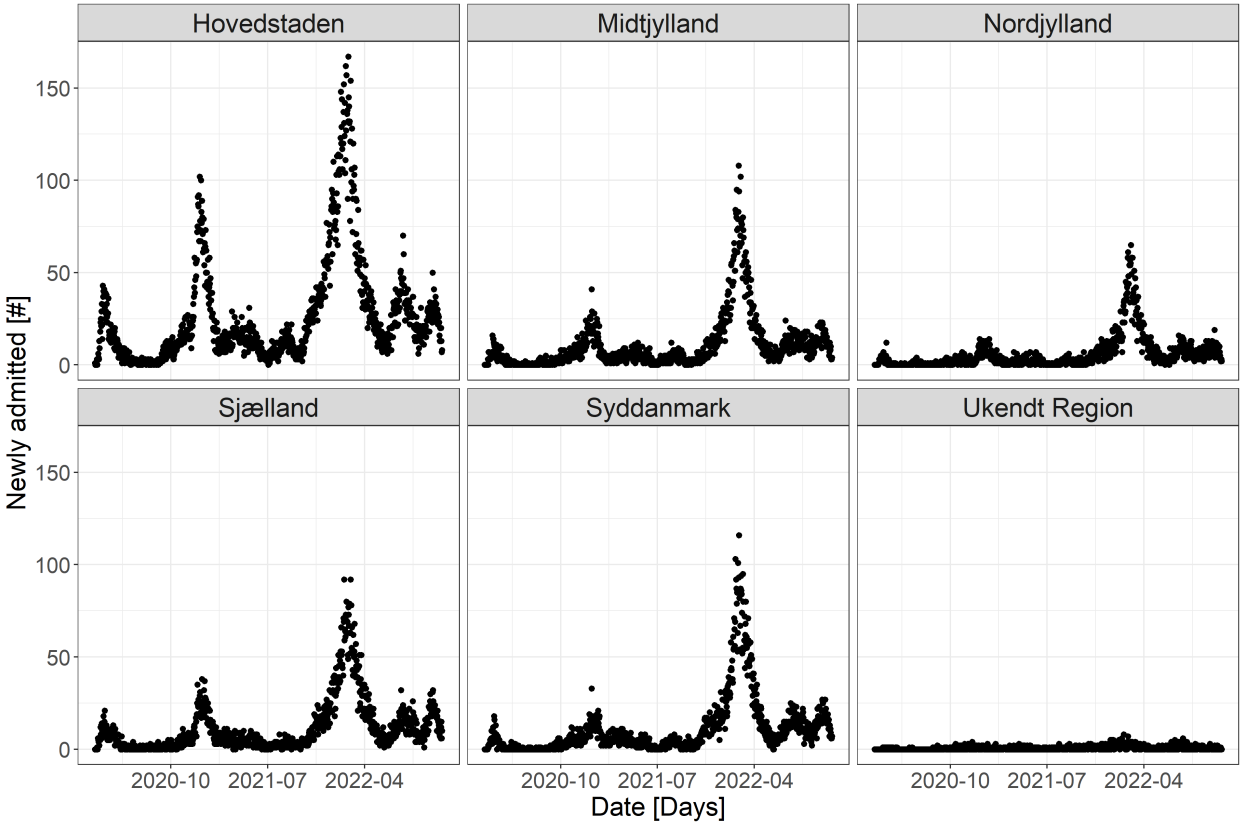

Figure 2: Total number of new hospital admissions in Denmark grouped by region of residence.

It can be seen that most of the new hospital admissions can be linked to *Hovedstaden*, while only a negligible amount are linked to *Ukendt Region*. Intermediate amount of new hospital admissions are linked to *Midtjylland*, *Nordjylland*, *Sjælland*, and *Syddanmark*. This is largely due to the fact, that *Hovedstaden* considers more individuals and that the exposure therefore is higher in this region. Additionally, a correlation between the regions are observed.

Methodology

In this section the GLMMs for modelling the number of new hospital admissions are formulated. Moreover, methods for approximating the likelihood functions and implementations in R packages for parameter estimation are presented.

Modelling

In order to analyze the data a simple state space model is proposed. The count observations h_t , $t = 1, \ldots, n$ in a period of $n = 981$ days starting on the 1st of March, 2020, of new hospital admissions is assumed to follow a Poisson distribution $h_t \sim \mathrm{P}(\lambda_t)$ with intensities given by

$$\log(\lambda_t) = \beta + u_t \tag{1}$$

Here β is a fixed effect parameter, that represents the average intensity and u_t is a random effect that is assumed to follow a first order auto-regressive process

$$u_t = au_{t-i} + \epsilon_t \tag{2}$$

where $\epsilon \sim N(0, \sigma^2), t > 1$ is a white noise process, and a and σ are model parameters. Using the results from Madsen (2007), it is assumed that u_1 follow the stationary distribution of the first order auto-regressive process $u_1 \sim N(0, \sigma^2/(1-a^2))$. Hence the joint likelihood becomes

$$L(a, \beta, \sigma; \mathbf{u}, \mathbf{y}) = \phi_{0, \frac{\sigma^2}{1-\sigma^2}}(u_1) \prod_{t=2}^{n} \left(\phi_{0,\sigma^2}(u_t - au_{t-1})\right) \prod_{t=1}^{n} \left(p_{\lambda_t}(h_t)\right) \tag{3}$$

where ϕ_{μ,σ^2} is the probability density function (pdf) of the normal distribution with mean μ and variance σ_2, and p_λ is the pdf of the Poisson distribution with mean λ. Intuitively, the model can be extended by modelling the individual regions. Hence, (a, β, σ) are 6-dimensional vectors.

In order to obtain the likelihood for the model parameters (a, β, σ) the observed random effects are integrated out. Hence, the marginal likelihood is obtained

$$L_M = (a, \beta, \sigma; \mathbf{y}) = \int_{\mathbb{R}^q} L(a, \beta, \sigma; \mathbf{u}, \mathbf{y}) d\mathbf{u} \tag{4}$$

where q is the number of random effects and a, β, and σ are the parameters to be estimated.

In order to make computation of the joint likelihood function in (4) feasible, the estimation is carried out using the multivariate Laplace approximation.

Laplace approximation

The marginal log-likelihood $l_M(a, \beta, \sigma; \mathbf{y}) = \log(L_M(a, \beta, \sigma; \mathbf{y}))$ is approximated by a second order Taylor approximation around the optimum $\tilde{\mathbf{u}} = \hat{\mathbf{u}}_\theta$ of the log-likelihood function w.r.t. the unobserved random variables $\mathbf{u}$, i.e.,

$$l(a, \beta, \sigma; \mathbf{u}, \mathbf{y}) \approx l(a, \beta, \sigma; \tilde{\mathbf{u}}, \mathbf{y}) - \frac{1}{2}(\mathbf{u} - \tilde{\mathbf{u}})^T H(\tilde{\mathbf{u}})(\mathbf{u} - \tilde{\mathbf{u}}) \tag{5}$$

where the first-order term of the Taylor expansion disappears since the expansion is done around the optimum $\tilde{\mathbf{u}}$ and $H(\tilde{\mathbf{u}}) = -l''_{uu}(a, \beta, \sigma, \mathbf{u}, \mathbf{y})|_{\mathbf{u}=\tilde{\mathbf{u}}}$ is the negative Hessian of the joint log-likelihood evaluated at $\tilde{\mathbf{u}}$.

Using the approximation in (5) on (4) the Laplace approximation of the marginal log-likelihood becomes (See Madsen & Thyregod (2011))

$$l_{M,LA}(a, \beta, \sigma; \mathbf{y}) = \log \int_{\mathbb{R}^q} \exp\left(l(a, \beta, \sigma; \tilde{\mathbf{u}}, \mathbf{y}) - \frac{1}{2}(\mathbf{u} - \tilde{\mathbf{u}})^T H(\tilde{\mathbf{u}})(\mathbf{u} - \tilde{\mathbf{u}})\right) d\mathbf{u} \tag{6}$$

$$= l(a, \beta, \sigma; \tilde{\mathbf{u}}, \mathbf{y}) - \frac{1}{2} \log \left| \frac{H(\tilde{\mathbf{u}})}{2\pi} \right| \tag{7}$$

Importance sampling

Importance sampling is a re-weighting technique for approximating integrals w.r.t. a density f by simulation in cases where it is not feasible to simulate from the distribution with density f. Instead it uses samples from a different distribution with density g, where the support of g includes the support of f.

Parameter estimation

In this section two R packages for used to estimate the parameters are presented. Namely, `glmmTMB` and `KFAS` which are available at CRAN.

glmmTMB

This section describes the R package `glmmTMB` by Brooks et al. (2017) for linear and GLMMs using Template Model Builder (TMB). The models are estimated using maximum likelihood estimation via TMB. Random effects are assumed to be Gaussian on the scale of the linear predictor and are integrated out using Laplace approximation. Additionally, gradients are calculated using automatic differentiation.

KFAS

This section goes into detail with the R package `KFAS` by Helske (2017) for state space modelling with observations from the exponential family. The `KFAS` package can perform Kalman filtering and smoothing with exact diffuse initialization using an univariate approach.

In `KFAS` the Poisson distribution with intensity λ_t and exposure term e_t together with the log-link is supported. Thus we have $\mathrm{E}\left(h_t|\log(\lambda_t)\right) = \mathrm{Var}\left(h_t|\log(\lambda_t)\right) = e_t\lambda_t$. In this report the exposure term is assumed to be constant, i.e. $e_t = 1$. Hence, the differences in time and between the regions are represented directly in the estimated parameters and latent state.

In order to make inferences of the Poisson model, `KFAS` finds a Gaussian model with the same conditional posterior mode as $\mathrm{P}(\lambda|\mathbf{h})$. This is done trough an iterative process with Laplace approximation of $\mathrm{P}(\lambda|\mathbf{h})$, where the updated estimates of $\log(\lambda_t)$ are computed via the Kalman filtering and smoothing from the approximating Gaussian model. The final estimates of $\log(\hat{\lambda}_t)$ correspond to the mode of $\mathrm{P}(\lambda|\mathbf{h})$. Generally, the difference between the mode and the mean is negligible. Nevertheless, our interest is focused on the intensity λ_t rather than the linear predictor $\log(\lambda_t)$.

Direct transformation from the linear predictor to the intensity introduces some bias. To solve this problem `KFAS` also contains methods based on importance sampling.

Results

In this section the obtained estimates for the fixed- and random effects are presented. Additionally, the model residuals are evaluated.

Estimating the total number of new hospital admissions in Denmark

The average intensity, β, estimates from `glmmTMB` and `KFAS` is listed in Table 2 together with the standard deviation of the random effect residuals, σ. It can be seen that the average intensity, β is estimated to be 3.175 (0.726) and 3.61 (0.028) using `glmmTMB` and `KFAS` respectively. It is noted that the standard deviation of the estimate from `glmmTMB` is rather high, and that the estimate from `KFAS` lies within one standard deviation from the `glmmTMB` estimate. The standard deviation of the estimate from `KFAS` is rather low, which is because the `KFAS` R package contains methods based on importance sampling.

The estimated average number of new hospital admissions considering only the fixed effect is 23.927 and 36.966. There is a noticeable difference between these two estimates of β, which comes down to the fact that `glmmTMB` and `KFAS` uses different methods for parameter estimation.

As an example, the likelihood computation in `KFAS` is an iterative procedure which is stopped using som stopping criteria, so the log-likelihood function contains some noise. This can affect the gradient computations in

methods like the Broyden-Fletcher-Goldfarb-Shanno (`BFGS`), which is used in this report, and can in theory give unreliable results. Therefore, using a derivative free method like Nelder-Mead can be recommended, even though it is not as computationally efficient.

Table 2: Parameter estimates for the GLMM given in (1) and (2) obtained using the `KFAS` and `glmmTMB` R packages. The standard deviation of the estimate is given in the parantheses. A method for calculating σ in non-Gaussian models is not supported in `KFAS`.

Parameter	glmmTMB	KFAS
β	3.175 (0.726)	3.61 (0.028)
σ	1.327	...

The smoothed estimate of the random effects, u_t, from `glmmTMB` and `KFAS` is visualized in Figure 3. It can be seen that the estimate of u_t is comparable using the two methods, but that `KFAS` consistently estimates u_t lower than `glmmTMB`. This is likely because `KFAS` estimated β to be a bit higher than `glmmTMB` and that the correlation between these two parameter estimates are rather high. Additionally, it is noted that as $u_t > 0$ the intensity is above the average intensity, and the number of new hospital admissions is above the estimated average, while as $u_t < 0$ the opposite is true. The signal is connected to the expected value via the log-link function. Thus the number of new hospital admissions, h_t, increases exponentially with u_t.

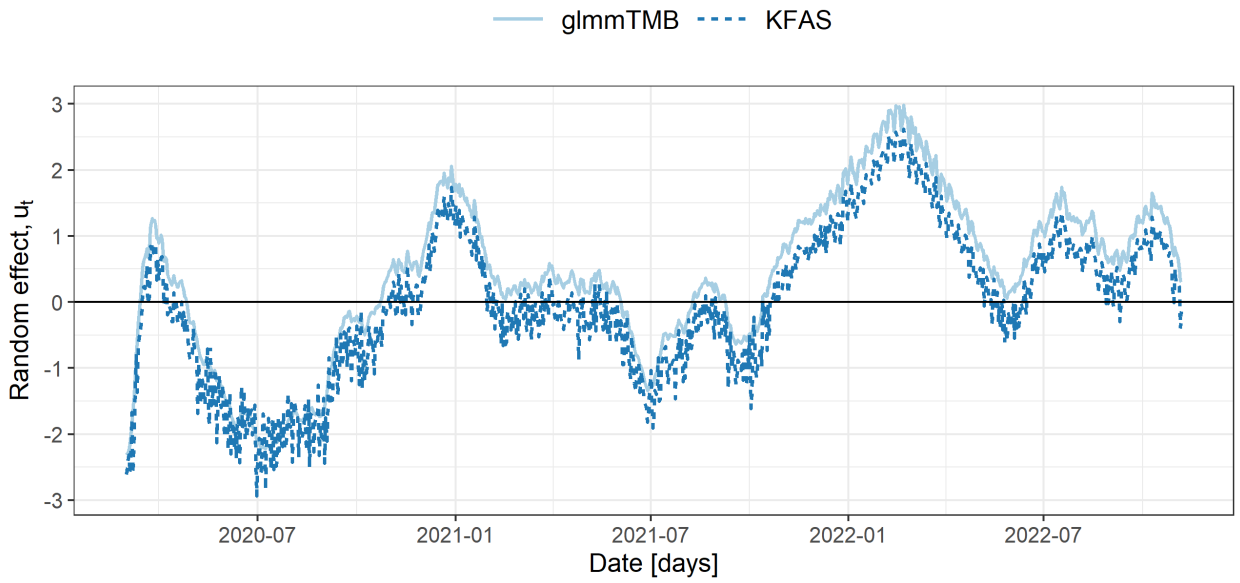

Figure 3: Smoothed estimates of the latent states from `glmmTMB` and `KFAS`.

In Figure 4 the smoothed estimates from `glmmTMB` and `KFAS` of the total number of new hospitals admissions in Denmark is visualized. There are no observable difference between the two implementations of the model.

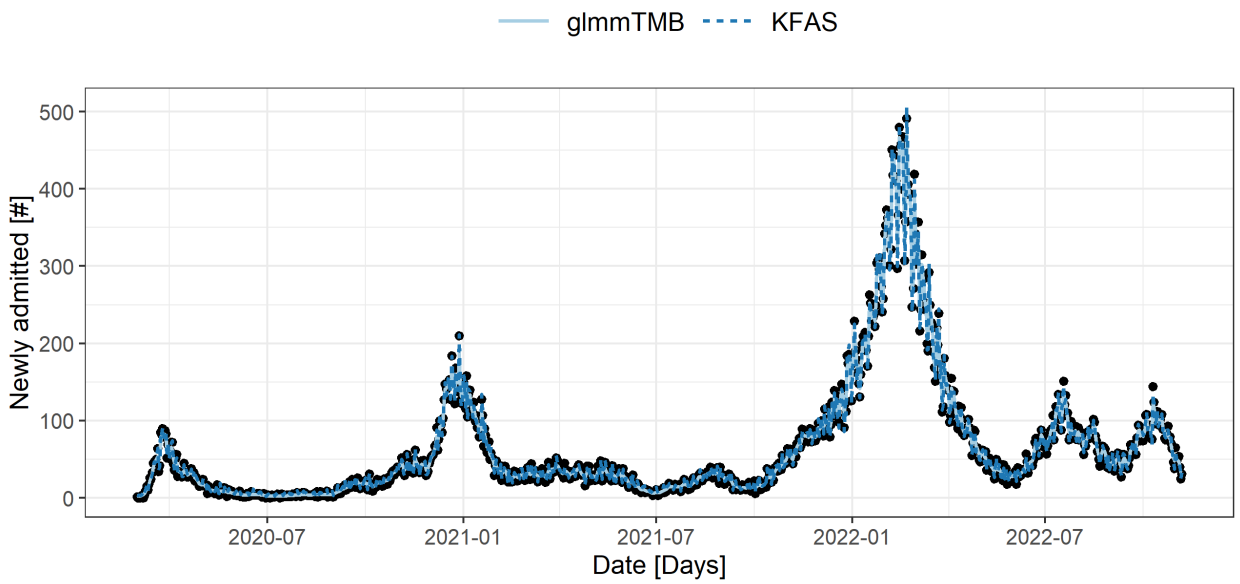

Figure 4: Smoothed estimates from `glmmTMB` and `KFAS` of the total number of new hospital admissions in Denmark.

Diagnostic plots of the one-step prediction residuals for the total number of new hospital admissions in Denmark are visualized in Figure 5. Overall the plots hint that the model can be improved. In A) and B) a significant auto-correlation in the residuals are detected. This is likely caused by the increased variance in the number of new admissions during the Covid-19 waves. The two lower plots, C) and D), indicate heavy tailed residuals.

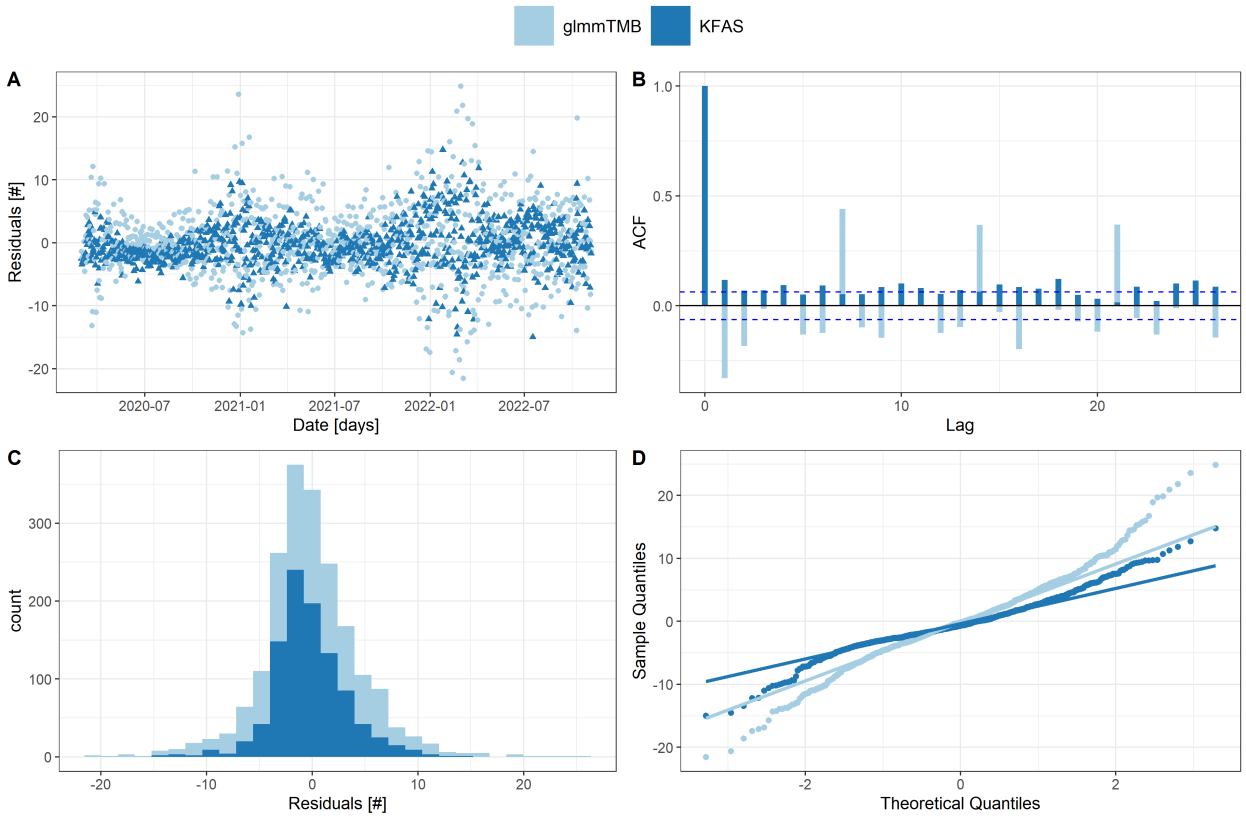

Figure 5: Diagnostic plots of the observed number of new hospital admissions in a specific regions minus the predicted number of new hospital admissions using `glmmTMB` and `KFAS`. A) Time series of the residuals, B) Auto-correlation function of the residuals, C) Histogram of the residuals, and D) Quantile-Quantile plot of the residuals.

Estimating the number of new hospital admissions in Denmark grouped by region

In this section the GLMM given in (1) and (2) is extended, such that the individual regions are modeled. Hence, the fixed effect parameter, β, is a 6-dimensional vector, $\beta^T = \left[\beta_{\text{Hov}}, \beta_{\text{Mid}}, \beta_{\text{Nor}}, \beta_{\text{Sjæ}}, \beta_{\text{Syd}}, \beta_{\text{Uke}}\right]$, the random effects vary independently for each region, u_t^r, $r = \left[\text{Hov, Mid, Nor, Sjæ, Syd, Uke}\right]$. The same white noise process, $e \sim N(0, \sigma^2)$, is assumed for all the regions.

The average intensity, β, estimates from `glmmTMB` and `KFAS` is gathered in Table 3 together with the standard deviation of the random effect residuals, σ. It can be seen that the average intensity in *Hovedstaden*, β_{Hov}, is estimated to be 2.438 and 2.63 using `glmmTMB` and `KFAS` respectively, which is the highest among the regions and corresponds to between 41.9-43.4% of of the average number of new hospital admissions. This is expected, as the exposure term is fixed to 1 in this report, and the difference in the population size between the regions is therefore observed within the parameter estimates.

As before, the uncertainty on the parameter estimates is higher using `glmmTMB` than `KFAS`, but this time some of the estimates for the average intensity lies within one standard deviation of each other for both `glmmTMB` and `KFAS`, and for β_{Nor} the maximum likelihood estimate is the same.

Table 3: Parameter estimates for the extended GLMM given in (1) and (2) obtained using the `KFAS` and `glmmTMB` R packages. The standard deviation of the estimate is given in the parantheses. A method for calculating σ in non-Gaussian models is not supported in `KFAS`.

Parameter	glmmTMB	KFAS
β_{Hov}	2.438 (0.612)	2.63 (0.017)
β_{Mid}	1.428 (0.611)	1.47 (0.095)
β_{Nor}	0.818 (0.612)	0.818 (0.088)
$\beta_{\text{Sjæ}}$	1.65 (0.61)	1.865 (0.156)
β_{Syd}	1.354 (0.611)	1.536 (0.145)
β_{Uke}	-0.971 (0.619)	-0.861 (0.018)
σ	1.245	...

In Figure 6 the random effects, u_t^r, $r = [\text{Hov, Mid, Nor, Sjæ, Syd, Uke}]$ is visualized. Obviously, a correlation between the regions is observed, which is expected as individuals residing in one region is not contained within that region and can spread disease to other regions as well. Hence, when an outbreak happens in one region it can leak into the adjacent regions.

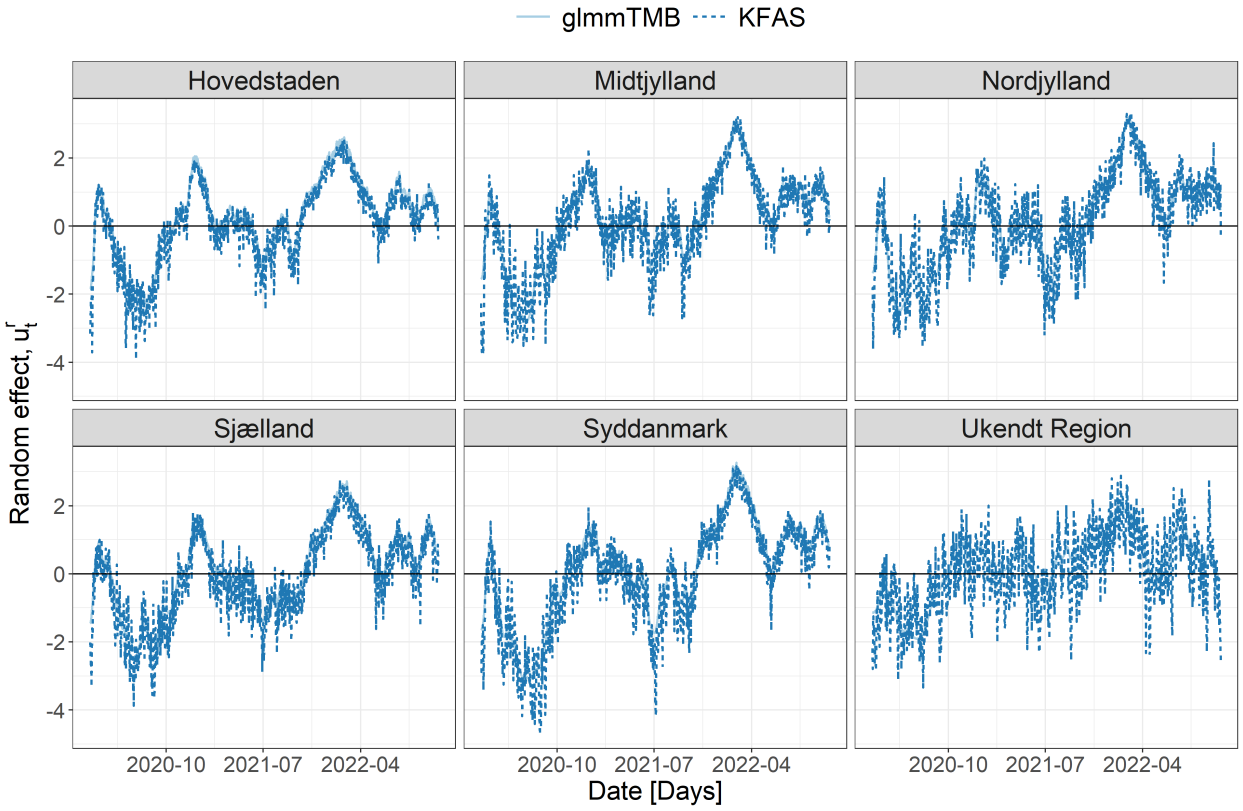

Figure 6: Smoothed estimates of the latent states from `glmmTMB` and `KFAS`.

In Figure 7 the smoothed estimate from `glmmTMB` and `KFAS` of the total number of new hospitals admissions in Denmark grouped by region is visualized. The difference between the two implementations are indistinguishable.

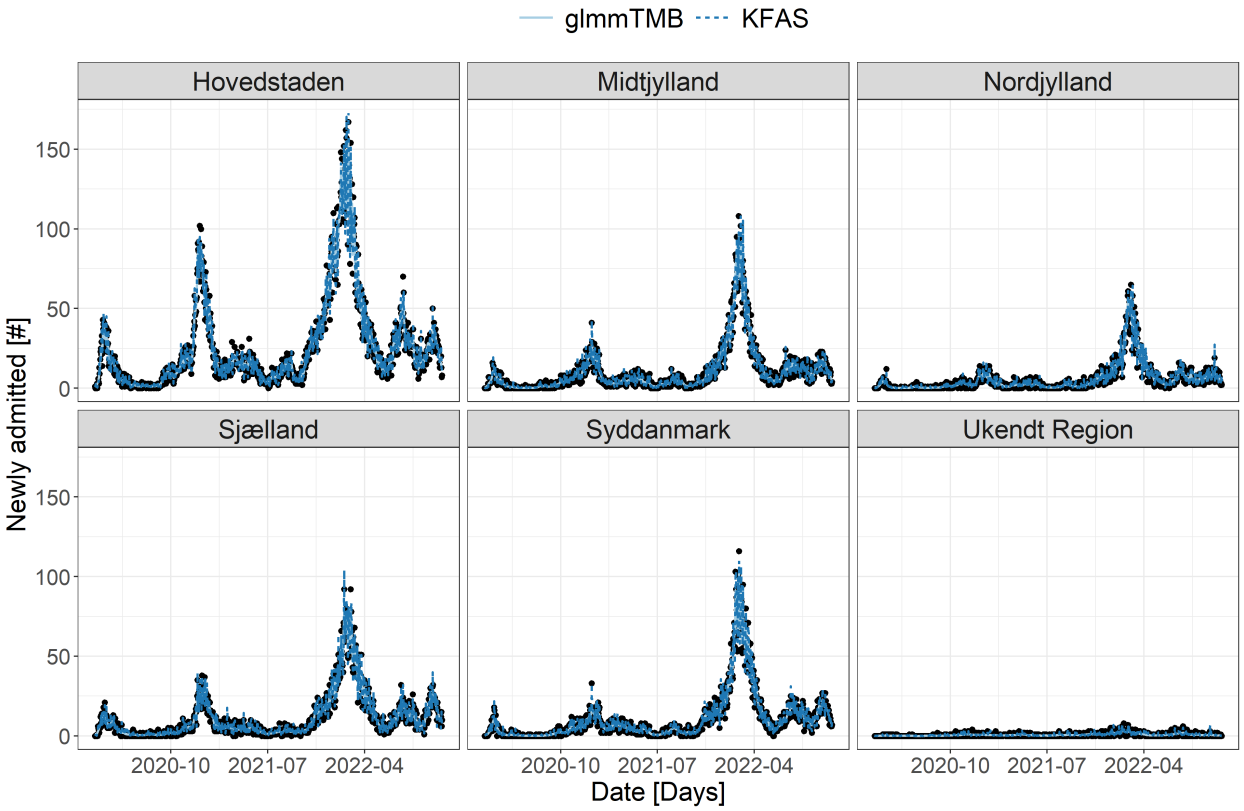

Figure 7: Smoothed estimates from `glmmTMB` and `KFAS` of the total number of new hospital admissions in Denmark grouped by region.

A time series of the residuals is visualized in Figure 8. It can be seen that the model is still not capable of describing the data sufficiently. The model struggles to predict the correct number of new hospital admissions when a wave of Covid-19 is occurring. Especially, this becomes apparent in *Hovedstaden*. Inspecting Figure 8 and Figure 11 from Appendix A, it can be seen that there are still a significant auto-correlation in the residuals. Furthermore, the Quantile-Quantile plot in Figure 10 in Appendix A indicates that the residuals are heavy-tailed. Hence, the model can be further improved.

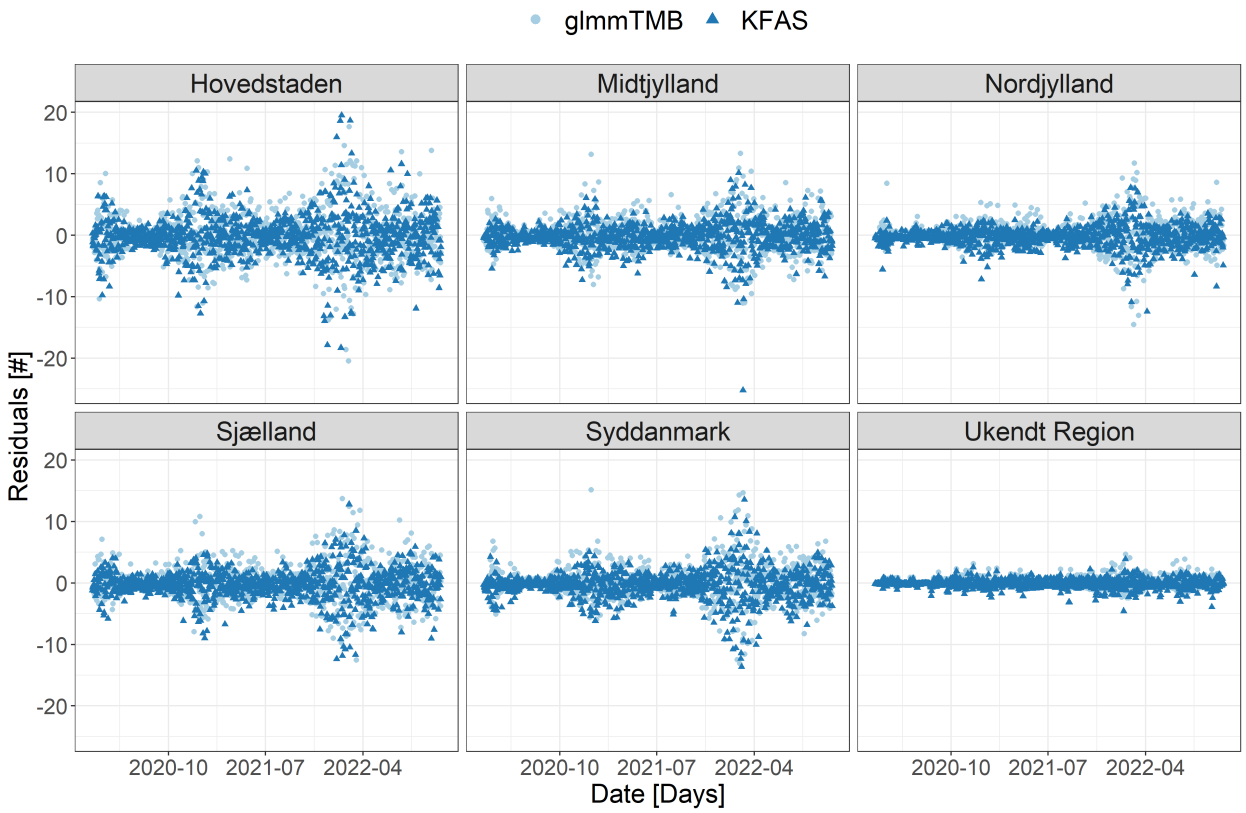

Figure 8: Time series of the observed number of new hospital admissions in a specific regions minus the predicted number of new hospital admissions using `glmmTMB` and `KFAS`.

Discussion

A generic model for early detection of disease outbreaks supported by data is in high demand.

As observed in Figure 2 and later confirmed in Figure 7 there is a clear correlation in the daily record of new hospital admissions in time and between the regions. Intuitively, this makes sense as infected individuals can move freely between the regions, and spread the disease. Henceforth, a possible extension to the model could be to include a correlation between the regions. Furthermore, the model could be extended with an extra white noise term in (1) in order to capture possible over-dispersion in the data.

While implementation of the GLMM in `glmmTMB` is rather straightforward, it is a bit more complicated in `KFAS`. On the other hand `KFAS` have a more flexible implementation, which allows for arbitrary model formulation by manually adjusting the system matrices. A major drawback for the `KFAS` implementation is that the computational costs of filtering and smoothing are relatively high, which makes it un-desired for online-filtering problems. Another possibility is to implement the GLMM using the R package Template Model Builder (`TMB`) by Kristensen et al. (2016), which is the package `glmmTMB` is build on. This would enable a more flexible implementation of the GLMM.

The R packages investigated in this report, `glmmTMB` and `KFAS`, are build on structurally different ideas and methods, which made it difficult to compare the implementations directly. However, it was found that both implementations sufficiently identified the random effects and could positively be used to detect a disease outbreak.

Conclusion

Clear auto-correlations are observed in the daily number of new hospital admissions. Additionally, a correlation between the regions are observed.

Both the `glmmTMB` and `KFAS` implementations sufficiently identifies the random effects and can be used to identify outbreaks with Covid-19 disease. Further extensions of the GLMM is needed in order to adequately describe the data.

References

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Helske, J. (2017). KFAS: Exponential family state space models in R. *Journal of Statistical Software*, *78*(10), 1–39. https://doi.org/10.18637/jss.v078.i10

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, *70*(5). https://doi.org/10.18637/jss.v070.i05

Madsen, H. (2007). *Time series analysis*. Chapman & Hall. https://doi.org/10.1201/9781420059687

Madsen, H., & Thyregod, P. (2011). *Introduction to general and generalized linear models*. CRC Press.

Appendix A

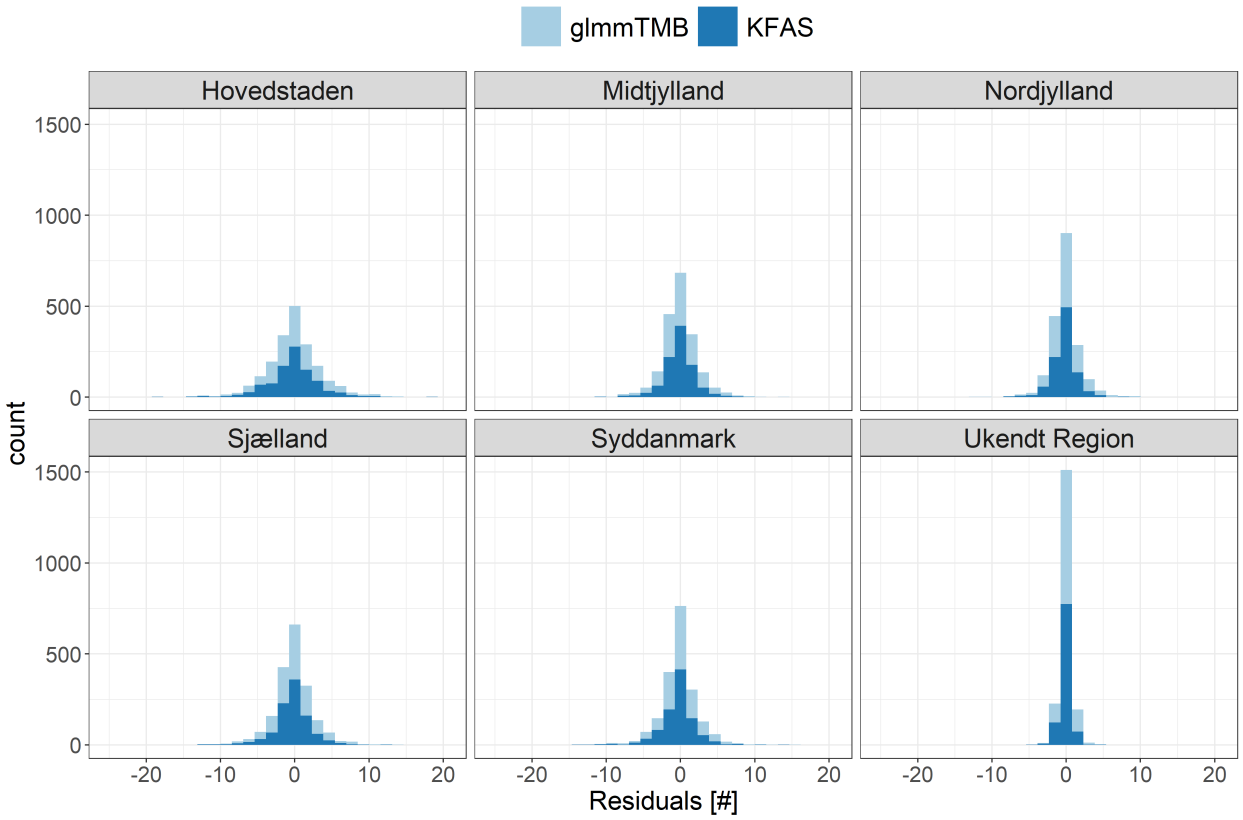

Figure 9: Time series of the observed number of new hospital admissions in a specific regions minus the predicted number of new hospital admissions using `glmmTMB` and `KFAS`.

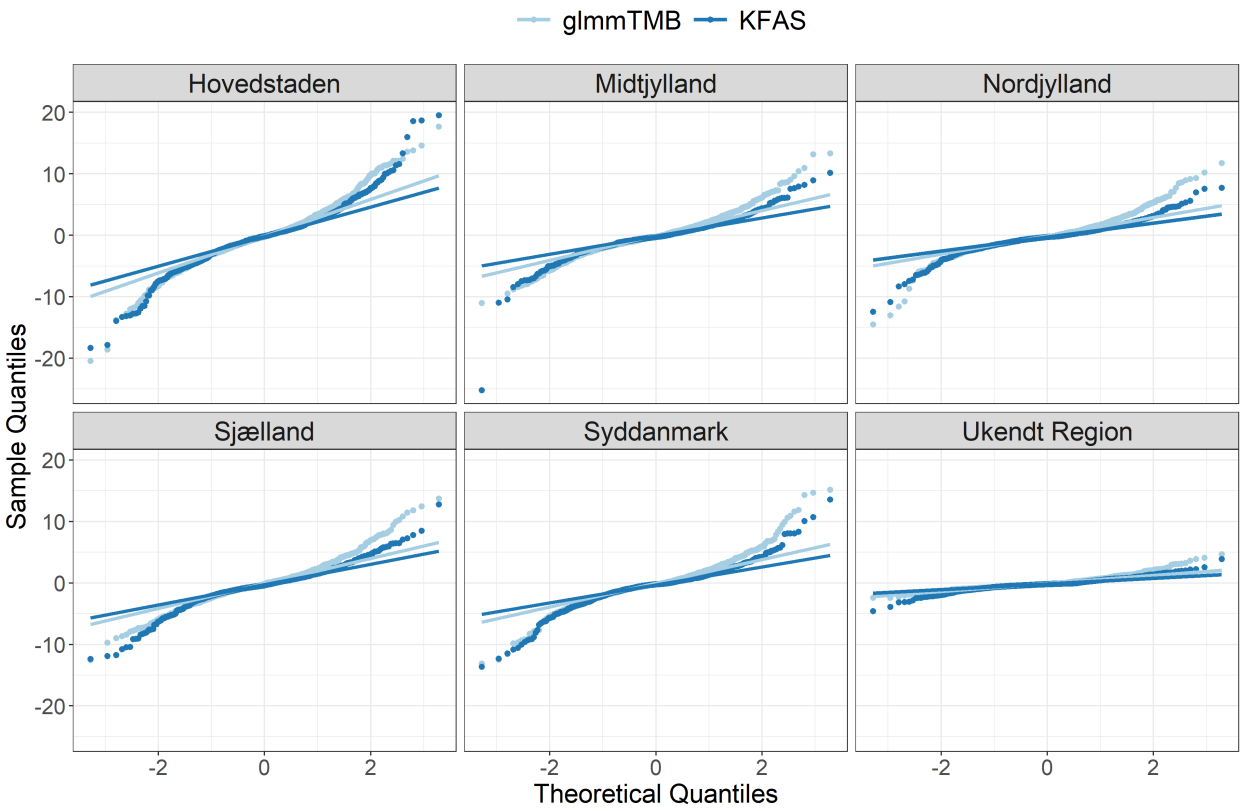

Figure 10: Quantile-Quantile plot of the observed number of new hospital admissions in a specific regions minus the predicted number of new hospital admissions using `glmmTMB` and `KFAS`.

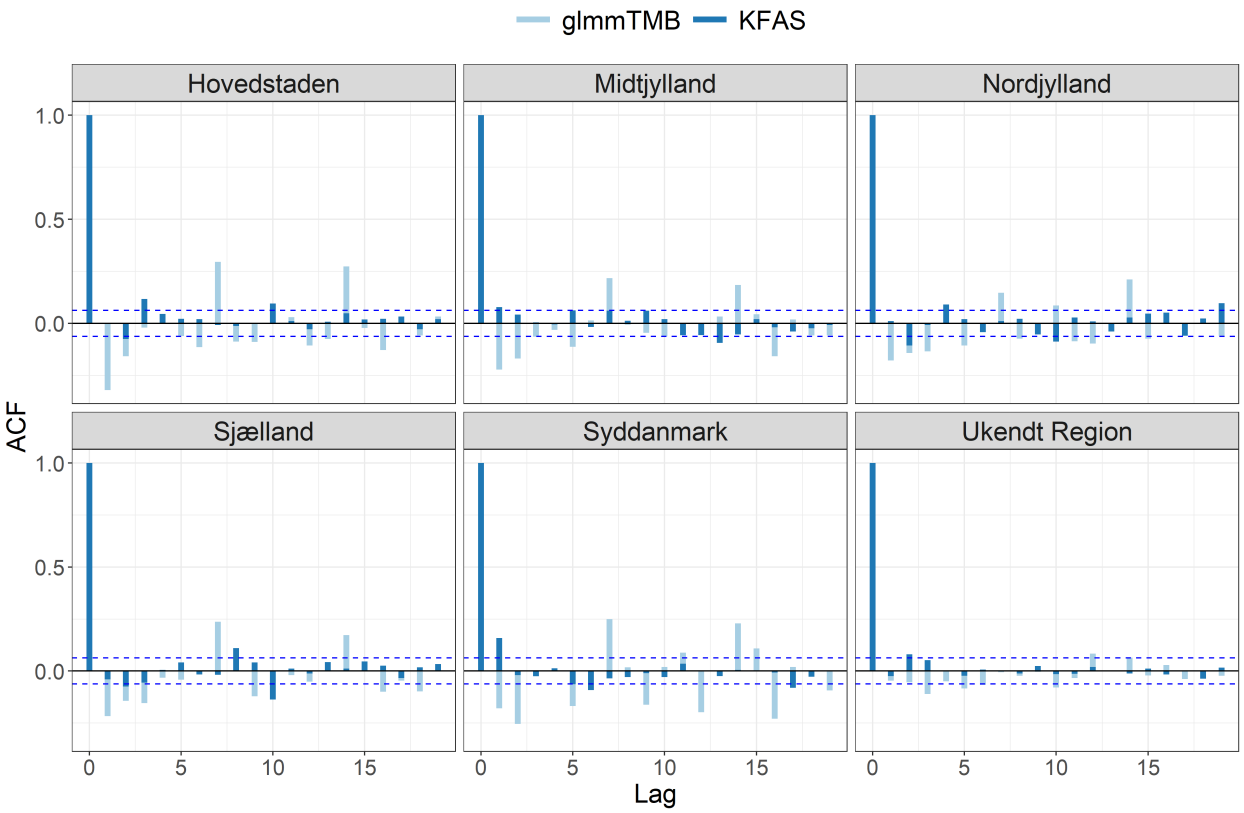

Figure 11: Auto-correlation function of the observed number of new hospital admissions in a specific regions minus the predicted number of new hospital admissions using `glmmTMB` and `KFAS`.