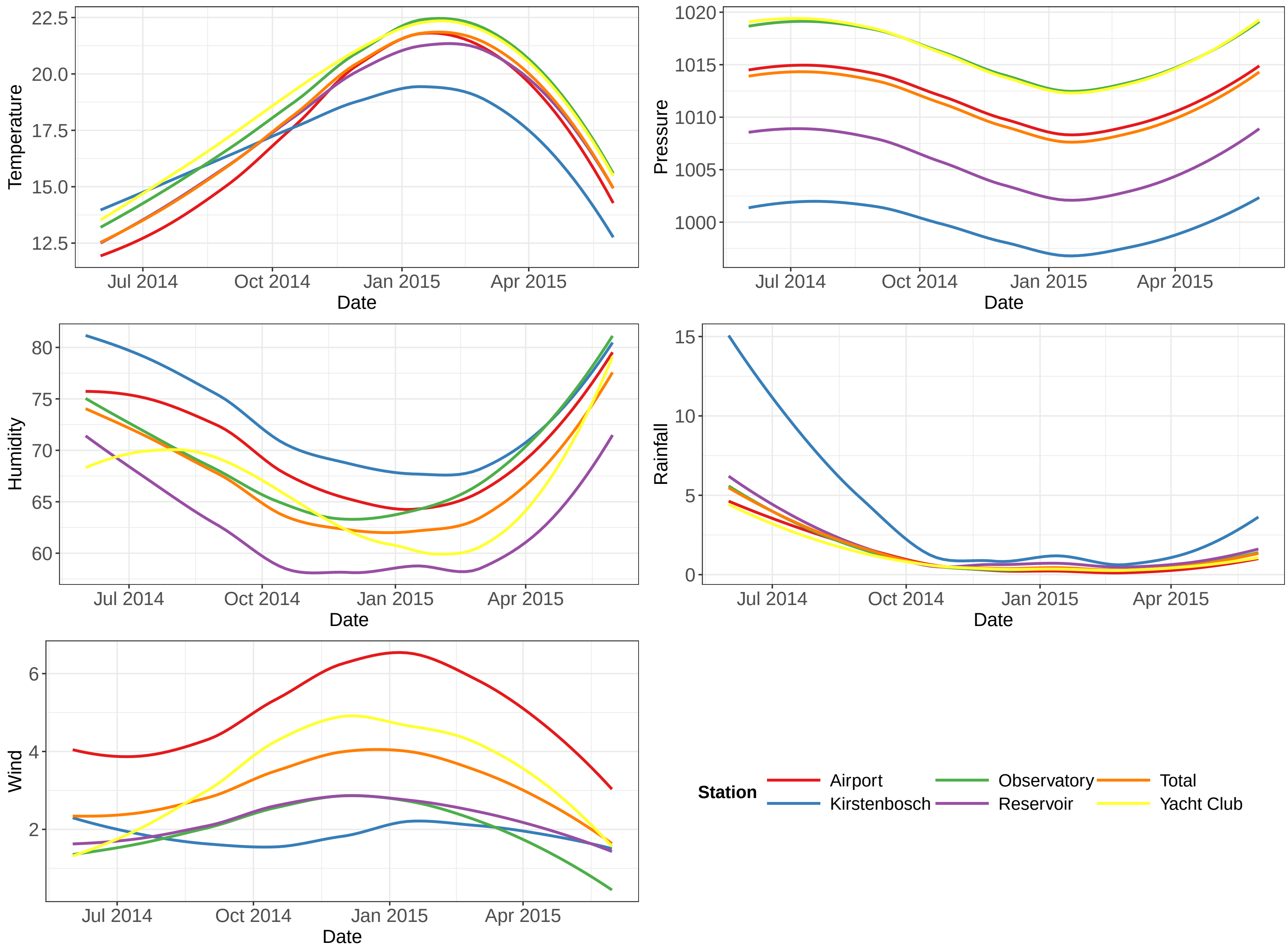
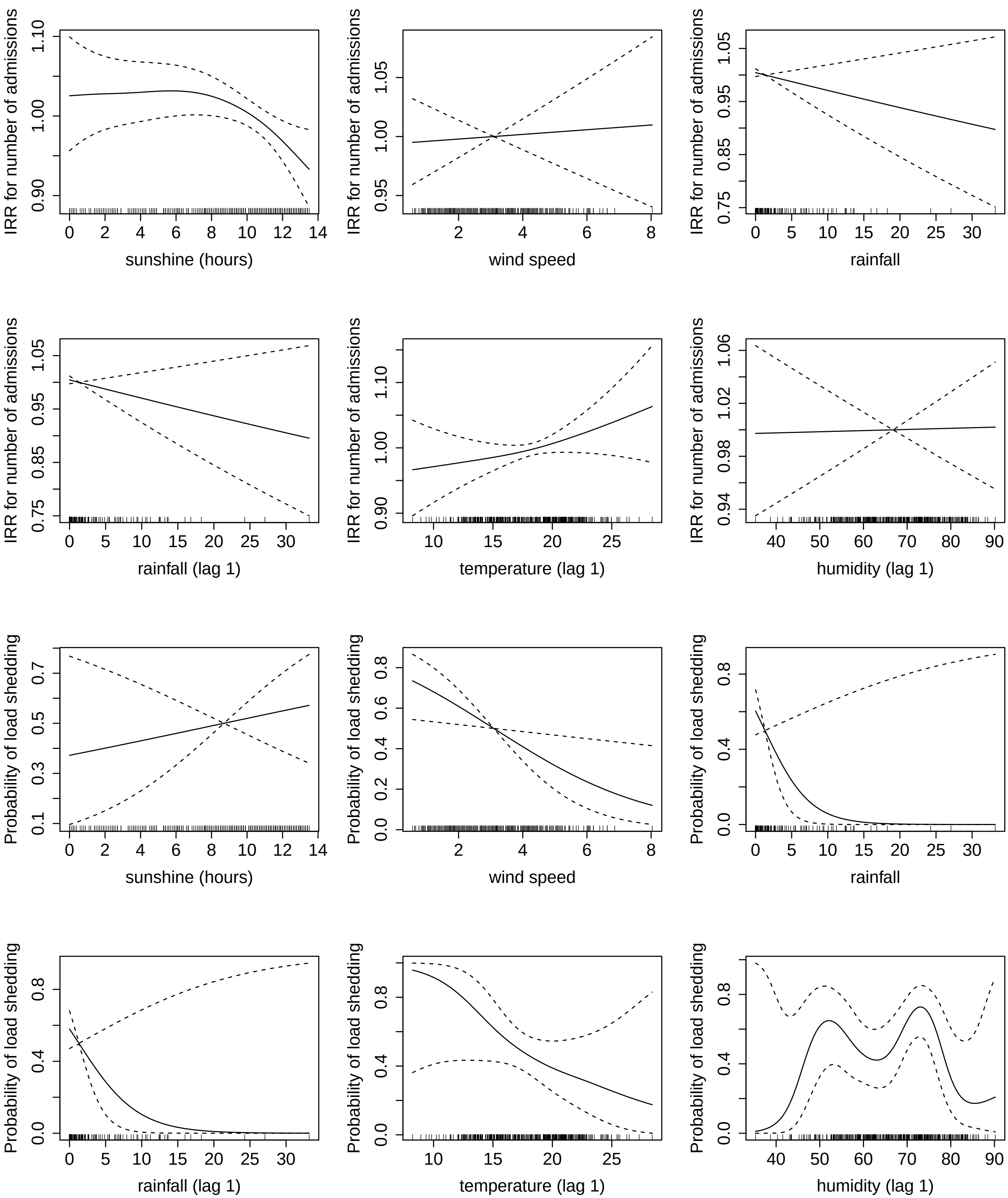
Supplementary Material for "Christian Gehringer, Heinz Rode, Michael Schomaker: *The effect of load shedding on pediatric hospital admissions in South Africa*"

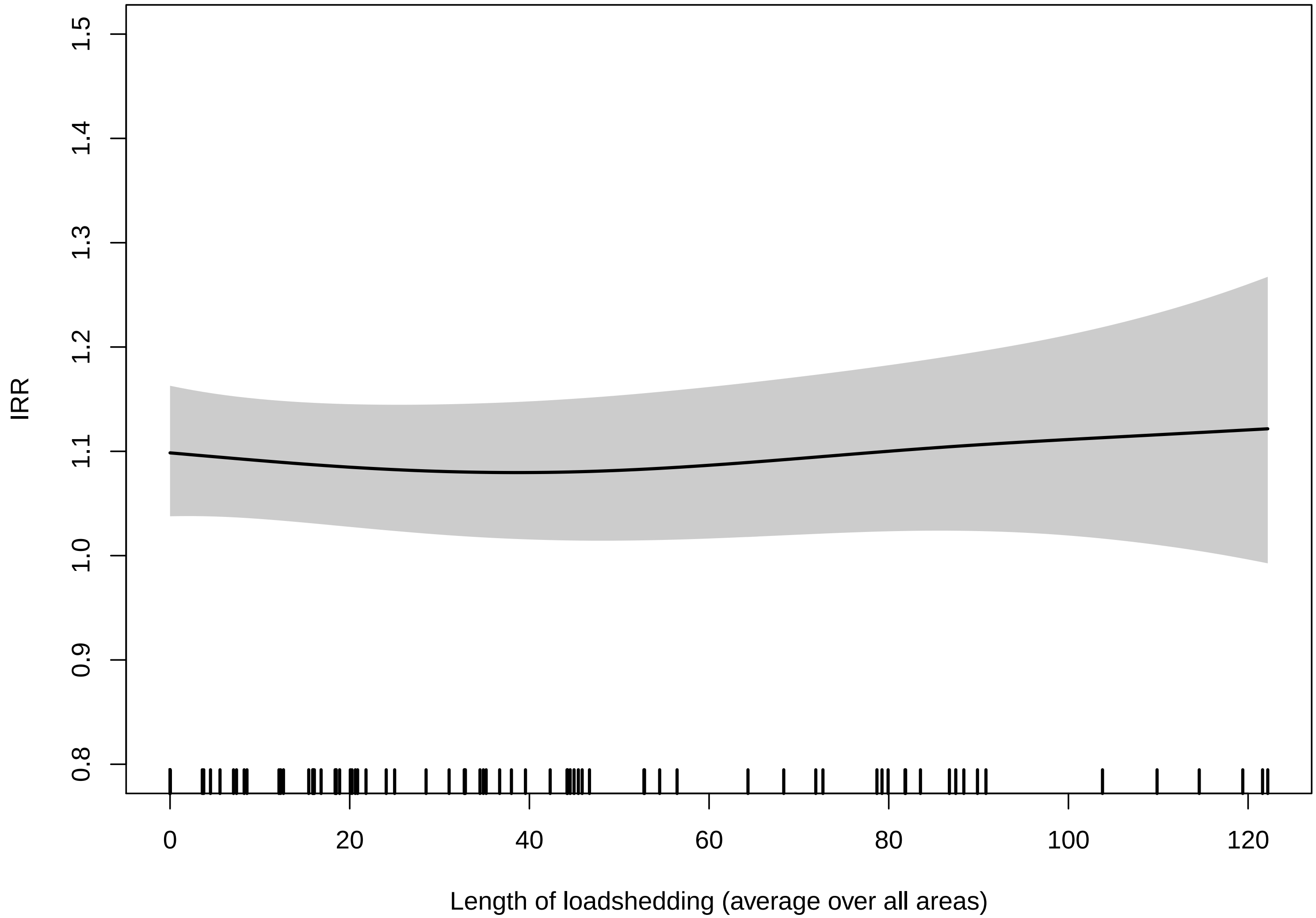
eFigure 1. Smoothed weather data for different weather stations in the Cape Town area: relative humidity (in %), pressure (in hectopascal), rainfall (in mm), temperature (in degrees Celsius), and wind speeds (in meter/second).



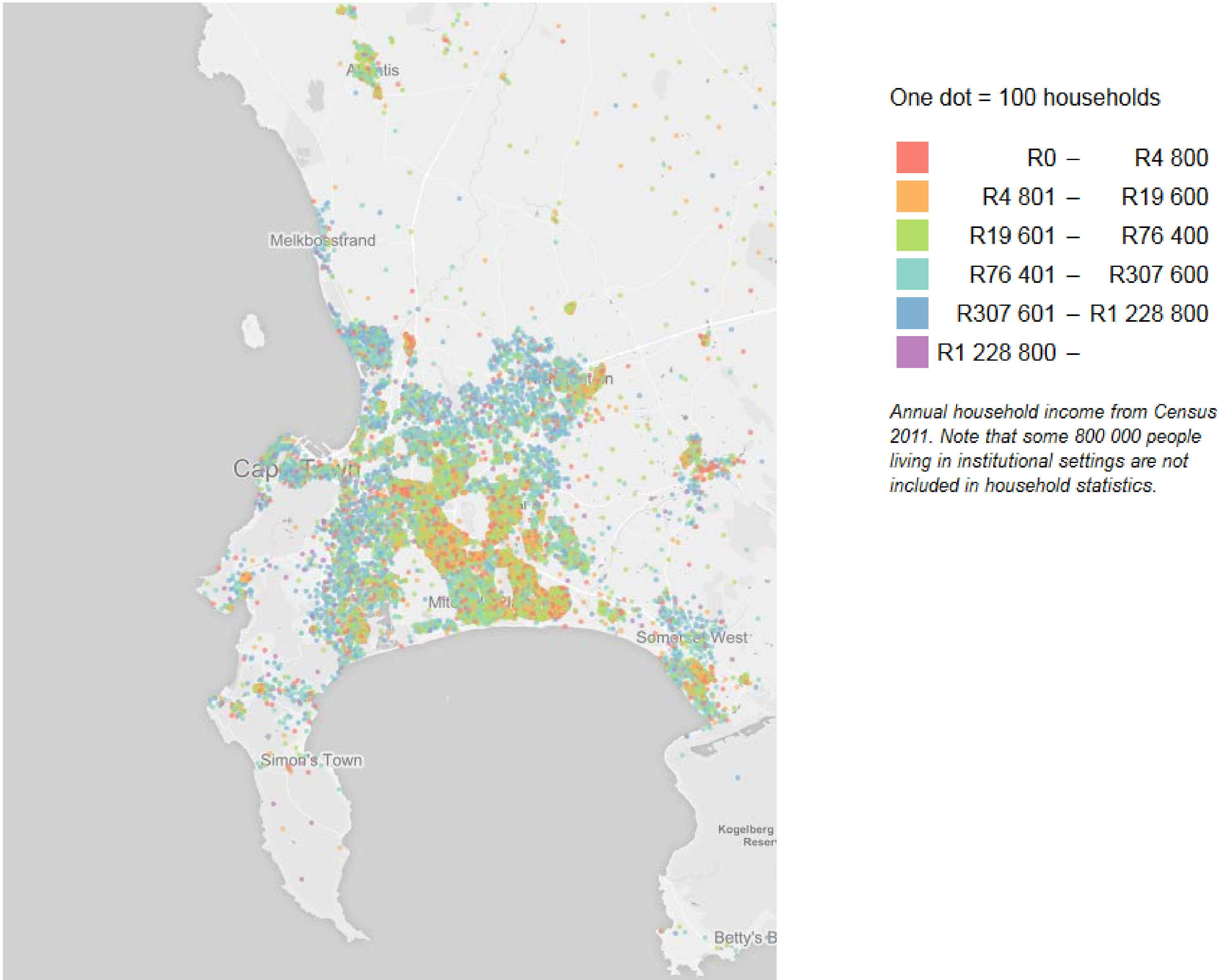
eFigure 2. Probability of load shedding (bottom 6 panels), and incidence rate ratios for number of hospital admissions (top 6 panels), depending on weather indicators calculated with an additive (covariate adjusted) logistic regression model. The reported weather indicators are sunshine (in hours), wind speed (in meter/second), rainfall (in mm), temperature (in degrees Celsius), and relative humidity (in %). Each panel shows the respective estimated penalized spline, conditional on other weather indicators and month.



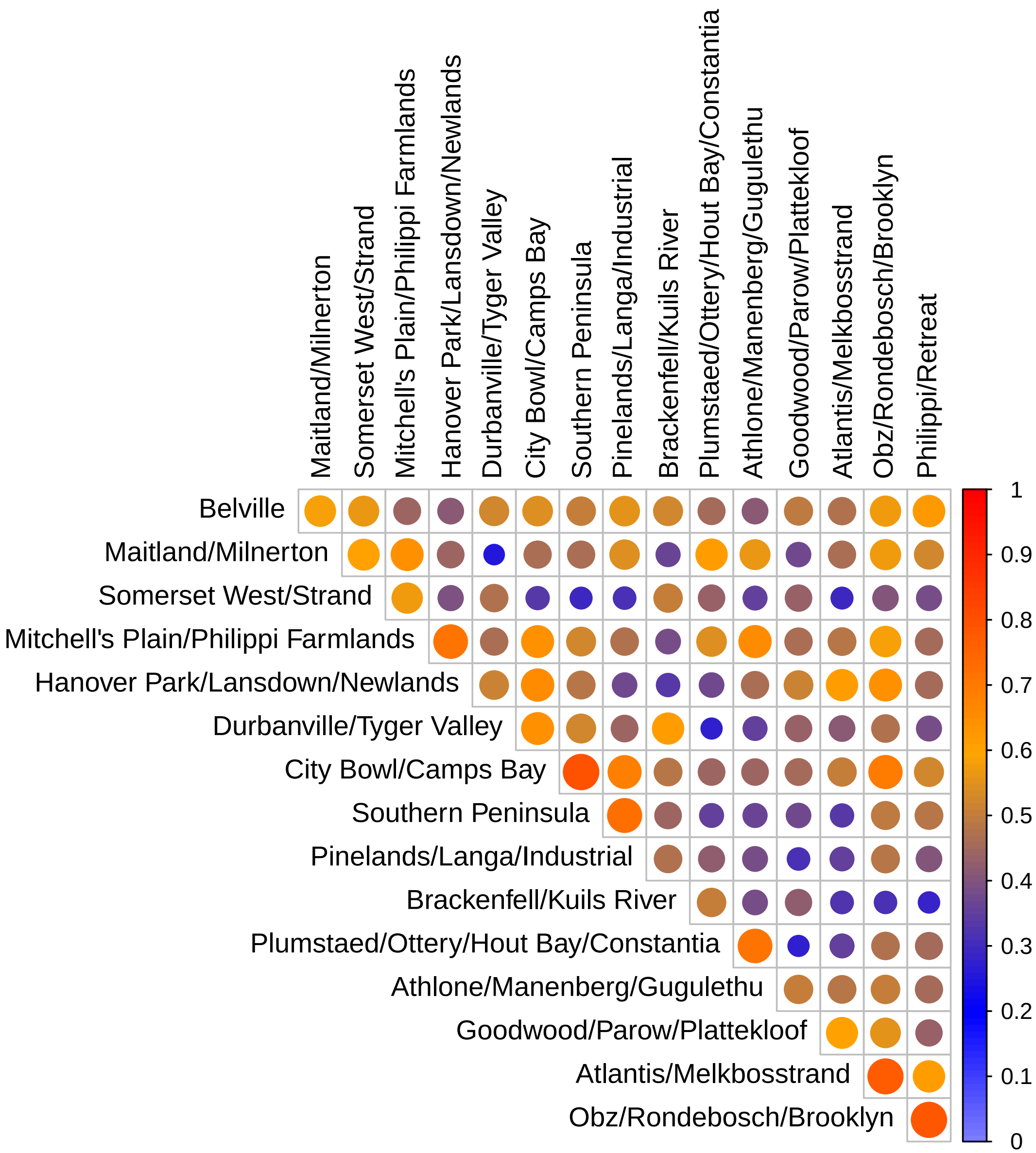
eFigure 3. Non-linear interaction between load shedding (same day or up to 2 days prior) and length of load shedding (average time, measured in minutes, per area and per day), modelled via penalized splines in an additive Quasi-Poisson model.



eFigure 4. Median yearly household income, as estimated in the national census from 2011, stratified by area. The dot map was produced by the tool of Adrian Firth, available at https://dotmap.adrianfrith.com/.



eFigure 5. Association between load shedding events in different areas measured by Cramer's V. The closer Cramer's V is to 1, i.e. the more red and the bigger the circle, and the stronger the association between two areas in terms of load shedding events taking place on the same day.



eTable 1. Average Treatment Effect (ATE) and incidence rate ratios (IRR) for diagnoses where the number of reported cases exceeds 25.

ICD-10	Diagnosis	IRR	95%	CI	ATE	95%	CI
J05	Acute obstructive laryngitis	1.74	0.53	5.73	0.05	0.01	0.10
T22	Burn & corrosion of shoulder & upper limb	1.47	0.45	4.85	0.29	0.21	0.37
T24	Burn & corrosion of lower limb	1.45	0.57	3.73	0.06	0.01	0.11
T23	Burn & corrosion of wrist and hand	1.37	0.49	3.83	0.01	-0.04	0.06
A41	Sepsis	1.34	0.62	2.91	0.00	-0.06	0.06
J44	Chronic obstructive pulmonary disease	1.29	0.58	2.88	-0.08	-0.16	-0.00
Q21	Congenital malformations of cardiac septa	1.27	0.37	4.34	0.61	0.49	0.73
J18	Pneumonia	1.24	0.74	2.07	0.00	-0.12	0.13
S09	Unspecified injuries of head	1.08	0.75	1.57	0.05	-0.08	0.18
G40	Epilepsy	1.08	0.54	2.16	-0.13	-0.23	-0.03
Z00	General examination	1.07	0.36	3.16	0.18	0.10	0.27
J21	Acute bronchiolitis	1.01	0.56	1.80	-0.24	-0.40	-0.07
T20	Burn & corrosion of head, face, and neck	0.95	0.42	2.14	-0.10	-0.17	-0.03
J45	Asthma	0.90	0.40	2.03	-0.07	-0.14	-0.01
T29	Burns of multiple regions	0.87	0.41	1.86	-0.13	-0.20	-0.06
A87	Viral meningitis	0.87	0.25	3.04	0.23	0.13	0.33
T74	Child abuse, neglect & other maltreatment	0.81	0.33	2.02	0.02	-0.02	0.07
A09	Infectious gastroenteritis	0.79	0.52	1.19	-0.12	-0.36	0.12
S42	Fracture of shoulder and upper arm	0.79	0.34	1.82	-0.04	-0.10	0.01
T21	Burn & corrosion of trunk	0.69	0.31	1.55	0.00	-0.07	0.08
S72	Fracture of femur	0.62	0.30	1.30	-0.09	-0.16	-0.02
S52	Fracture of forearm	0.61	0.30	1.23	-0.12	-0.19	-0.05
S06	Intracranial injury	0.59	0.16	2.17	-0.07	-0.11	-0.03
S82	Fracture of lower leg	0.59	0.25	1.40	-0.10	-0.15	-0.04
G03	Meningitis	0.56	0.13	2.50	-0.05	-0.10	-0.01
L02	Cutaneous abscess	0.52	0.24	1.15	-0.10	-0.18	-0.02
J22	Acute lower respiratory infection	0.37	0.14	0.99	-0.11	-0.17	-0.04

eText 1. Technical details about the statistical approaches used in the paper.

Content:

Notation and Variables. We are interested in the number of admissions at Red Cross hospital Y_i, measured on days $i = 1\ldots, 365$. Then, $E(Y)$ refers to the expected number of admissions per day. The (main) interventions of interest are the binary variable A indicating whether load shedding has been implemented on day i or not, and $\bar{A}^2$ which denotes whether load shedding occurred on the same or any of the two preceding days. The matrix of potential confounders $\mathbf{L}$ includes hours of sunshine (L^1), wind speed (L^2), humidity (L^3), pressure (L^4), precipitation (L^5), temperature (L^6), month (L_7), last week of the month (week of payment, L_8), and a seasonal (weekly) trend modeled via sine and cosine terms, i.e.

$$\cos(\omega_k t) \quad \text{and} \quad \sin(\omega_k t) \quad \text{with} \quad \omega_k = \frac{2k\pi}{T}$$

with $T = 7$ days. More specifically we consider $\cos(\omega_1 t) = L^9$, $\sin(\omega_1 t) = L^{10}$, $\cos(\omega_2 t) = L^{11}$, $\sin(\omega_2 t) = L^{12}$, $\cos(\omega_3 t) = L^{13}$, $\sin(\omega_3 t) = L^{14}$, $\cos(\omega_4 t) = L^{15}$, $\sin(\omega_4 t) = L^{16}$ to be potential confounders. Moreover, in most of the below described models, we also consider past values of the weather indicators up to a lag of 2 days as well as past admissions (i.e. $Y_{i-\text{lag}}$, $\text{lag} \in \{1, 2, \ldots, 13, 14, 21, 28\}$) to be of interest.

Frequentist Model Averaging. Since the potential set of adjustment variables is large, mostly due to the number of lagged and seasonal variables, we need to reduce the number of variables, so that the implementation of the below specified models is computationally feasible at all (i.e. for successful fitting of the splines and the causal inference procedures). A classic approach would be model selection with hypothesis testing, or using a criterion such as Akaike's Information Criterion (AIC). However, data-driven model selection has the disadvantage that model selection uncertainty is not being taken into account, i.e. different samples of data may yield different conclusions with respect to variable inclusion (Chatfield, 1995). An alternative to model selection is model averaging.

With model averaging, one calculates a weighted average $\hat{\bar{\beta}} = \sum_\kappa w_\kappa \hat{\beta}_\kappa$ from the k estimators $\hat{\beta}_\kappa$ ($\kappa = 1, \ldots, k$) of the set of candidate (regression) models $\mathcal{M}$ where the weights are calculated in a way such that 'better' models receive a higher weight. A popular weight choice would be based on the exponential AIC,

$$w_\kappa^{\text{AIC}} \quad = \quad \frac{\exp(-\frac{1}{2}\text{AIC}_\kappa)}{\sum_{\kappa=1}^{k} \exp(-\frac{1}{2}\text{AIC}_\kappa)}\,,$$

where AIC_κ is the AIC value related to model $M_\kappa \in \mathcal{M}$ (Buckland et al., 1997). Note that the weights sum up to one. The k point and k variance estimates can then be combined into a single model, based on the weights. In our analysis we are not interested in this combined estimate but rather into a variable importance measure obtained from the model averaging weights. With this, the sum of the weights of those models which include the variable of interest are simply being added up. So, if there is a single best model (which has by far the lowest AIC), then this model receives weight 1 and all variables contained in it receive a variable importance measure of 1 (and the others 0). However, if there are say 10 very good models (measured by AIC), and only 4 of those contain the respective variable, then the sum of the weights of those 4 models would determine the variable's importance. We use a 50% support, i.e. a variable importance of 0.5 or higher, as a rule to include a variable into our final model. More details are given below. Readers who are interested into model averaging may consult the following references: Burnham and Anderson (2002), Hjort and Claeskens (2003), Hoeting et al. (1999).

Quasi-Poisson Model: Model Specification. Our considerations start with a Poisson model which considers all potential variables of interest:

$$E(Y) \quad = \quad \exp(\beta_0 + \beta_1 A + \mathbf{L}\boldsymbol{\beta_2}) \tag{1}$$

where $\mathbf{L}$ includes $L^1, \dots, L^{16}$, $L^1_{i-1}, \dots, L^6_{i-1}$, $L^1_{i-2}, \dots, L^6_{i-2}$, and Y_{i-1}, $1 \in \{1, 2, \dots, 13, 14, 21, 28\}$. Note that a standard Poisson model assumes $E(Y) = Var(Y) = \mu$ which may not be met (and in fact is not met in our data[1]). For this reason we use below only Quasi-Poisson models which allow $Var(Y) = \phi E(Y)$, where ϕ is a parameter which is constant across all $i = 1, \dots, n$. The parameter ϕ can be estimated from the data using the χ^2-statistic, and can also be used to correct the standard maximum likelihood estimate $\widehat{\text{Var}}(\hat{\beta}_{i,ML})$, i.e. one can use $\hat{\phi} \times \widehat{\text{Var}}(\hat{\beta}_{i,ML})$ to construct confidence intervals for $\hat{\beta}_i$. With this approach we still use the model equation (1), but with more flexible variance estimation (McCullagh and Nelder, 1989).

To reduce the number of variables we consider different models in which a number of variables are being held fixed and others vary (i.e. a group of related variables, say lagged variables) so that model averaging and the above introduced variable importance measure can be calculated[2].

i) To determine the lags needed for $Y_{i-\text{lag}}$ we implement model averaging based on the following full Quasi-Poisson model:

$$\begin{aligned} E(Y) \quad = \quad & \exp(\beta_0 + \beta_1 \bar{A}^2 + \beta_2 Y_{i-1} + \beta_3 Y_{i-2} + \beta_4 Y_{i-3} + \beta_5 Y_{i-4} + \beta_6 Y_{i-5} + \beta_7 Y_{i-6} + \beta_8 Y_{i-7} + \beta_9 Y_{i-8} + \beta_{10} Y_{i-9} \\ & + \beta_{11} Y_{i-10} + \beta_{12} Y_{i-11} + \beta_{13} Y_{i-12} + \beta_{14} Y_{i-13} + \beta_{15} Y_{i-14} + \beta_{16} Y_{i-21} + \beta_{17} Y_{i-28} + \beta_{18} L^9 + \beta_{19} L^{10} + \beta_{20} L^{11} \\ & + \beta_{21} L^{12} + \beta 22 L^7 + \beta_{23} L^8 + \beta_{24} L^1 + \beta_{24} L^2 + \beta_{24} L^5 + \beta_{25} L^3_{i-1} + \beta_{26} L^5_{i-1} + \beta_{27} L^6_{i-1} + \beta_{28} L^2_{i-2}) \end{aligned}$$

This analysis yields variable importance measures > 0.5 for the following lags of Y: $1, 3, 7, 9$.

ii) To determine the amplitude of the seasonal component, i.e. the k needed for $\cos(\omega_k t)$ and $\sin(\omega_k t)$, we implement model averaging based on the following full Quasi-Poisson model:

$$\begin{aligned} E(Y) \quad = \quad & \exp(\beta_0 + \beta_1 \bar{A}^2 + \beta_2 Y_{i-1} + \beta_3 Y_{i-3} + \beta_4 Y_{i-7} + \beta_5 Y_{i-9} \\ & + \beta 6 L^7 + \beta_7 L^8 + \beta_8 L^1 + \beta_9 L^2 + \beta_{10} L^5 + \beta_{11} L^3_{i-1} + \beta_{12} L^5_{i-1} + \beta_{12} L^6_{i-1} + \beta_{14} L^2_{i-2} \\ & \beta_{15} L^9 + \beta_{16} L^{10} + \beta_{17} L^{11} + \beta_{18} L^{12} + \beta_{19} L^{13} + \beta_{20} L^{14} + \beta_{21} L^{15} + \beta_{22} L^{16}) \end{aligned}$$

Cosine and sine terms (i.e. $L^9 - L^{16}$) that are supported by a variable importance of > 0.5 are $k = 1$ and $k = 2$.

iii) To determine the importance of the potential confounder weather we implement model averaging for the following two Quasi-Poisson and logistic regression models:

$$\begin{aligned} E(Y) \quad = \quad & \exp(\beta_0 + \beta_1 \bar{A}^2 + \beta_2 Y_{i-1} + \beta_3 Y_{i-3} + \beta_4 Y_{i-7} + \beta_5 Y_{i-9} + \beta 6 L^7 + \beta_7 L^8 + \beta_8 L^9 + \beta_9 L^{10} + \beta_{10} L^{11} + \beta_{11} L^{12} \\ & \beta_{12} L^1 + \beta_{13} L^2 + \beta_{14} L^3 + \beta_{15} L^4 + \beta_{16} L^5 + \beta_{17} L^6 + \beta_{18} L^1_{i-1} + \beta_{19} L^2_{i-1} + \beta_{20} L^3_{i-1} + \beta_{21} L^4_{i-1} + \beta_{22} L^5_{i-1} \\ & + \beta_{23} L^6_{i-1} + \beta_{24} L^1_{i-2} + \beta_{25} L^2_{i-2} + \beta_{26} L^5_{i-2} + \beta_{27} L^6_{i-2}) \end{aligned}$$

$$\begin{aligned} P(A = 1) \quad = \quad & \beta_0 + \beta_1 A_{i-1} + \beta_2 A_{i-2} + \beta 3 L^7 + \beta_4 L^8 + \beta_5 L^1 + \beta_6 L^2 + \beta_7 L^3 + \beta_8 L^4 + \beta_9 L^5 + \beta_{10} L^6 + \beta_{11} L^1_{i-1} \\ & + \beta_{12} L^2_{i-1} + \beta_{13} L^3_{i-1} + \beta_{14} L^4_{i-1} + \beta_{15} L^5_{i-1} + \beta_{16} L^6_{i-1} + \beta_{17} L^1_{i-2} + \beta_{18} L^2_{i-2} + \beta_{19} L^3_{i-2} + \beta_{20} L^4_{i-2} \\ & + \beta_{21} L^5_{i-2} + \beta_{22} L^6_{i-2} \end{aligned}$$

Variables with an importance > 0.5 in either of the two models are $L^1, L^2, L^5, L^3_{i-1}, L^5_{i-1}, L^6_{i-1}, L^2_{i-2}$.

The above considerations yield to the following final (additive) Quasi-Poisson model which has been used to obtain the main results in Table 1:

$$\begin{aligned} E(Y) \quad = \quad & \exp(\beta_0 + \beta_1 \bar{A}^2 + \mathbf{L}^* \boldsymbol{\beta}_2 + f_1(Y_{i-1}) + f_2(Y_{i-3}) + f_3(Y_{i-7}) + f_4(Y_{i-9}) + f_5(L^1) + f_6(L^2) + f_7(L^5) \\ & + f_8(L^3_{i-1}) + f_9(L^5_{i-1}) + f_{10}(L^6_{i-1}) + f_{11}(L^2_{i-2})) \end{aligned} \tag{2}$$

In the above equation, $\mathbf{L}^* = (L^7, L^8, L^9, L^{10}, L^{11}, L^{12})$. With $f(\cdot)$ we refer to unspecified smooth functions which we fit with penalized splines, as implemented in the R-package `mgcv` (Wood, 2017). The estimate $\exp(\hat{\beta}_1)$ is the one reported in Table 1.

[1] We have overdispersion, i.e. the ratio of the deviance to the degrees of freedom is about 1.4. The estimated dispersion parameter for quasi poisson family is 1.39935.

[2] with R-package `MuMIn` (Barton, 2017)

Targeted Maximum Likelihood Estimation (in our analysis). From a causal perspective we are interested in the counterfactual outcome $Y_i^{A=a}$ which refers to the hypothetical outcome that would have been observed if at day i there had been, possibly contrary to the fact, the intervention $A = a$, i.e. load shedding or not. More generally we would like to know the expected number of admissions per day for a particular intervention, $E(Y^a)$. Since A is binary, a sensible target quantity is

$$\psi \quad = \quad E(Y^1) - E(Y^0),$$

that is the average treatment effect meaning that we are interested in the difference in expected number of admissions per day had load shedding being implemented during the whole year compared to if this had not been the case. If we look at $\bar{A}^2$ rather than A we can interpret ψ as the difference in expected number of admissions per day had there been a load shedding event each day or on any of the preceding two days, during the whole year, compared to if this had not been the case. This causal parameter can be identified under the following assumptions:

1. Consistency: if A is binary, then $Y_i = A_i Y_i^1 + (1 - A_i)Y_i^0$.
2. Positivity: $P(A = a|L = l) > 0$ for $\forall l$ with $P(L = l) \neq 0$.
3. Conditional exchangeability: $Y^a \coprod A|L$ for $\forall A = a, L = l$.

Consistency can be interpreted as having a well-defined intervention, that can't be interpreted/implemented in multiple ways. This assumption could be met in our data as implementation of load shedding as a temporary power-shutdown is well-defined in the sense that the consequence of load shedding is the unavailability of electricity. Positivity requires a positive probability of treatment assignment in all confounder strata, i.e. a positive probability of load shedding occurrence no matter what season and weather. There are no practical or theoretical considerations which would point towards violation of this assumption, however there may be practical positivity violations in the sample data because of the small sample size. An indication of possible positivity violations would be very small ($<< 0.01$) estimated probabilities of treatment assignment used in the TMLE procedure (see below). In our data probabilities varied between 0.01781 and 0.89374 showing no signs of severe practical positivity violations. Conditional exchangeability would be violated if there are unmeasured confounders. Our DAG (Figure 1, main manuscript) explains why we believe we have measured the main confounders, though we can't exclude the possibility of unmeasured confounders. Moreover, we consider the seasonal trend to be modeled correctly; if this is not the case both conditional exchangeability and the (conditional) independence assumption needed for the modeling (i.e. the validity of the likelihood functions used and application of super learning, see below) could be violated and this may introduce bias.

The theory and application of TMLE has been described elsewhere[3]. Briefly, TMLE requires the fitting of both the conditional expectation of the outcome $E(Y|A, \mathbf{L})$ and the treatment mechanism $P(A = 1|\mathbf{L})$. Modeling the conditional outcome enables standardization with respect to the confounders, i.e. integrating $\mathbf{L}$ out, which equates to the (parametric) g-formula:

$$E(Y^a) \quad = \quad \int_l E(Y|A = a, L = l)\, dF_l(l)$$

where $F_L(\cdot)$ is the cumulative distribution function with respect to L. Thus, the average treatment effect ψ can be obtained by using the g-formula to calculate $E(Y^1)$ and $E(Y^0)$. Practically this equates to fitting an appropriate regression model and predicting the outcome for the whole sample for both $A = 1$ and $A = 0$[4]. TMLE adds an additional *targeted* step whereby the initial estimate is updated by a regression model which contains i) a fixed intercept/offset, which is the initial estimate of the ATE (from the g-formula) and ii) a clever covariate which is a function of the inverse probabilities of treatment assignment (based on $P(A = 1|\mathbf{L})$). This update reduces bias (if present) and improves efficiency (if no bias present). TMLE can readily incorporate machine learning to estimate $E(Y|A, \mathbf{L})$ and $P(A = 1|\mathbf{L})$ while retaining valid inference. We have implemented TMLE with super learning, a combination of machine learning and statistical forecasting.

Super learning means considering a set of prediction algorithms, for example regression models, shrinkage estimators or boosting. Instead of choosing the algorithm with the smallest cross validation error, super learning chooses a weighted

[3]Good introductions for epidemiologists can be found in Luque Fernandez et al. (2018) and Schuler and Rose (2017)

[4]The empirical distribution, i.e. the data, is taken for $L = l$

combination of different algorithms, that is the weighted combination which minimizes the cross validation error. It can be shown that this weighted combination will perform (asymptotically) at least as good as the best algorithm, if not better (Van der Laan et al., 2008) and this is known as the oracle property of super learning. Briefly, super learning works as follows:

1. First split the data into blocks of equal size (i.e. ten blocks of 100 observations for a sample size of 1,000 units and the choice of 10-fold cross-validation) and fit each of the selected algorithms on the training set (i.e. on 9 out of the 10 blocks).

2. Then, predict the estimated probabilities of the outcome (Y) using the validation set (i.e. the remaining one block) for each algorithm.

3. Repeat steps 1 and 2 for each of the ten blocks. This yields predictions for all 1,000 observations for each learning algorithm.

4. Now, estimate the cross validated risk for each learning algorithm, that is a function of the true values of Y and the respective predictions, for example the (vector of the) squared differences.

5. Then, use non-negative least squares estimation to find the weighted linear combination of cross validated risks (related to each learner) which predicts Y best. Note that the weights sum up to one.

6. Then, use the weights to create a weighted prediction from the different learning algorithms. This yields the super learner estimate of $E(Y|A, \mathbf{L})$ and $P(A = 1|\mathbf{L})$ respectively.

In our analysis we have implemented TMLE with the following specifications [for the main result in Table 1]:

- Our outcome Y is the number of hospital admissions, $A = \bar{A}^2$ denotes whether load shedding occurred on the same or any of the two preceding days, and the considered confounders are (based on the model building considerations above, i.e. variable reduction with model averaging for computational feasibility) $L^1, L^2, L^5, L^3_{i-1}, L^5_{i-1}, L^6_{i-1}$, $L^2_{i-2}, L^7, L^8, L^9, L^{10}, L^{11}, L^{12}, Y_{i-1}, Y_{i-3}, Y_{i-7}, Y_{i-9}$.

- Our target quantity is ψ, as discussed above.

- We use the *R*-package `tmle` to implement super learning (Gruber and van der Laan, 2012).

- We use super learning to estimate $E(Y|A, \mathbf{L})$ and $P(A = 1|\mathbf{L})$ with 10-fold cross validation, a squared loss function (for cross validation), and a truncation level of 0.01, meaning that estimated probabilities $P(A = 1|\mathbf{L})$ (needed for the targeted update step) would have been truncated if they were lower than 0.01 (though this did not occur). We use the following learners to estimate $E(Y|A, \mathbf{L})$: a full linear regression model, a linear regression model with stepwise AIC based model selection, a linear regression model with AIC based forward selection, GLM's based on an EM-algorithm-Bayesian model fitting (Gelman and Su, 2016), the arithmetic mean, a linear regression model with stepwise AIC based model selection including interaction terms, LASSO estimation (Tibsharani, 1996), LASSO averaging (Schomaker, 2012) for the full model and models with interactions and squared terms (Schomaker, 2017), boosting (Ridgeway, 2017), multivariate adaptive regression splines (Milborrow, 2017), Mallow's model averaging (Hansen, 2007), Jackknife Model Averaging (Hansen and Racine, 2012), linear regression after screening variables with LASSO, multivariate adaptive regression splines after screening variables with LASSO, and boosting after screening with Cramer's V (Heumann et al., 2016, Chapter 4). The same learners have been used to estimate $P(A = 1|\mathbf{L})$ except Mallow's Model Averaging, Jackknifel Model Averaging, and the LASSO. In addition k-nearest neighbour classification has been used.

The Negative Binomial Model. Recall the Quasi-Poisson model used to obtain the results reported in Table 1:

$$\begin{aligned} E(Y) \quad &= \quad \exp(\beta_0 + \beta_1 \bar{A}^2 + \mathbf{L}^* \boldsymbol{\beta}_2 + f_1(Y_{i-1}) + f_2(Y_{i-3}) + f_3(Y_{i-7}) + f_4(Y_{i-9}) + f_5(L^1) + f_6(L^2) + f_7(L^5) \\ &\quad\quad + f_8(L^3_{i-1}) + f_9(L^5_{i-1}) + f_{10}(L^6_{i-1}) + f_{11}(L^2_{i-2})) \end{aligned}$$

An alternative to using the above model, would still be using the above model, but instead of using the assumption that the observations Y_i follow a Poisson distribution (conditional on the covariates), one can assume that they follow a negative binomial distribution:

$$P(Y = y) \quad = \quad \frac{\Gamma(\phi + y)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda}\right)^{\phi} \left(\frac{\lambda}{\phi + \lambda}\right)^{y}$$

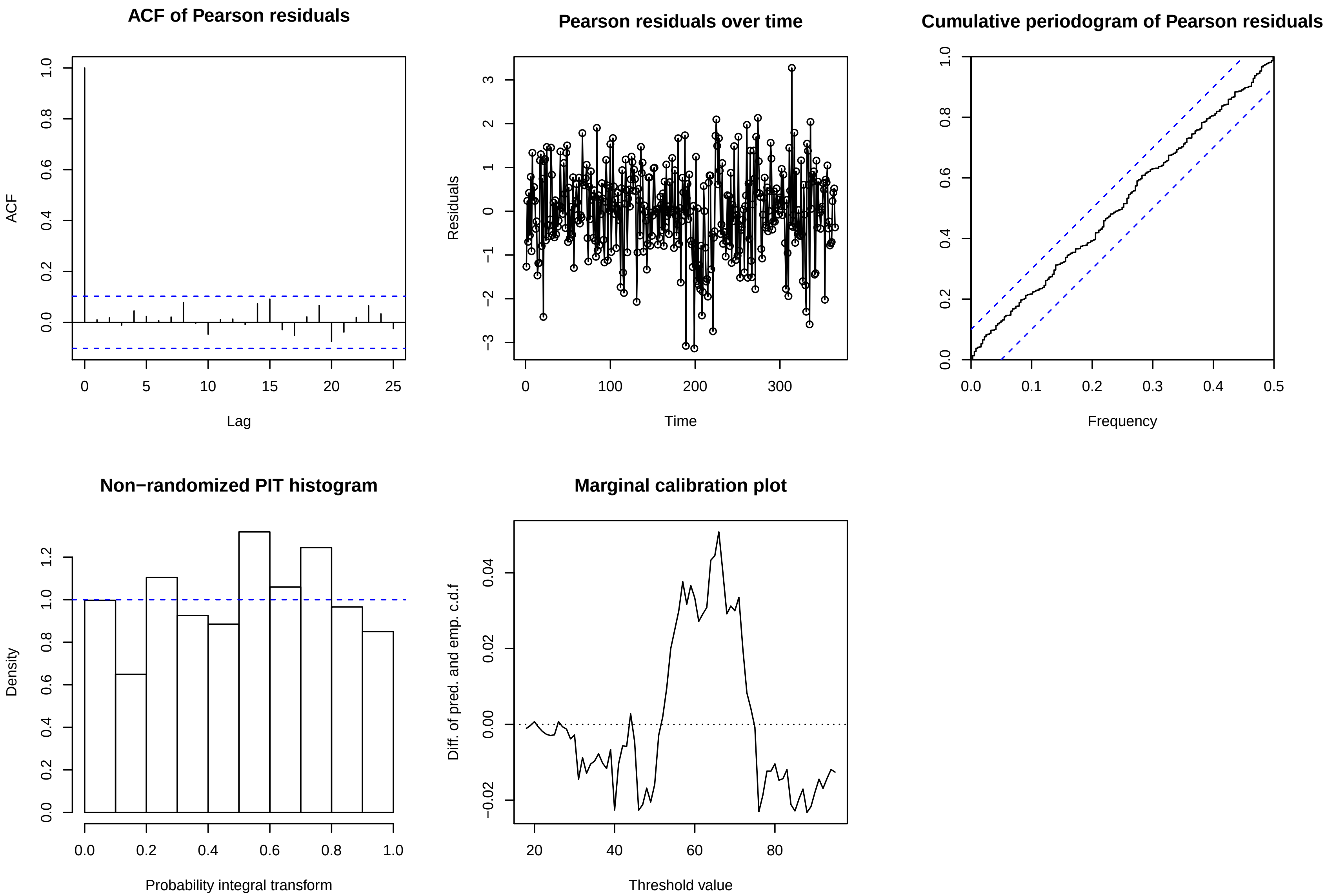
Note that $\Gamma(n)$ is the Gamma function, defined as $\Gamma(n) = (n-1)!$ for positive integers and $\Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)dt$ otherwise. For $\phi \to \infty$ this distribution approaches a Poisson distribution. While for a Poisson distribution we get a variance of λ, it is $\lambda + \lambda^2/\phi$ for the negative binomial approach, thus allowing to adjust for overdispersion. Our results are based on the negative binomial model as explained above, implemented with the `glm.nb` from the library `MASS`. No splines have been used to include the covariates.

The INGARCH model. The integer-valued GARCH model, also known as the autoregressive Poisson model, is similar to the models used above – with the difference that the mean conditional on the past ($E(Y_i|\mathcal{F}_i)$) is modeled, i.e. we explicitly acknowledge the time-series structure by modeling a stochastic process conditional on the past. With $\mathcal{F}_i$ we mean the history of the joint process $\{Y_i, A_i, \mathbf{L}_i\}$. In our analysis, we use the following INGARCH model for which we assume a negative-binomial distribution of Y_i:

$$E(Y) \quad = \quad \exp(\beta_0 + \alpha_1\lambda_{i-1} + \alpha_2\lambda_{i-2} + \alpha_3\lambda_{i-3} + \beta_1\bar{A}^2 + \tilde{\mathbf{L}}\boldsymbol{\beta}_2)$$

where $\tilde{\mathbf{L}}$ contains all variables from equation (2) and $\lambda_i = E(Y_i|\mathcal{F}_i)$. The decision to include the conditional means up to a lag of 3 has been based on model selection with AIC. Thus, the INGARCH model is identical to our negative-binomial model above, with the exception of the inclusion of three conditional means. Background on inference of these models, as well as good references, can be found in Liboschik et al. (2017). This reference also explains useful diagnostic for the INGARCH model, all of which are shown in eFigure 6. The autocorrelation function of the residuals does not exhibit any serial correlation that hasn't been taken into account by the model (as the autocorrelation is small and below the blue limits). There is also no sign that the residuals are inappropriate. However, the probability integral transform (PIT) histogram is not ideal, it should be more uniformly distributed. Moreover, the marginal calibration plot, which plots the difference between the average predictive cumulative distribution function (cdf) and the empirical cdf, shows major deviations from zero. This means that the number of hospital admissions, given the past, is not modelled well by the INGARCH model. It is also worth mentioning that with the INGARCH approach we haven't been able to include non-linear relationships with splines.

eFigure 6. Diagnostics for the fitted INGARCH model.



References

Barton, K. (2017). *MuMIn: Multi-Model Inference*. R package version 1.40.1/r411.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997). Model selection: an integral part of inference. *Biometrics 53*, 603–618.

Burnham, K. and D. Anderson (2002). *Model selection and multimodel inference. A practical information-theoretic approach*. Springer, New York.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society A 158*, 419–466.

Gelman, A. and Y.-S. Su (2016). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. R package version 1.9-3.

Gruber, S. and M. J. van der Laan (2012). tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software 51*(13), 1–35.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica 75*, 1175–1189.

Hansen, B. E. and J. Racine (2012). Jackknife model averaging. *Journal of Econometrics 167*, 38–46.

Heumann, C., M. Schomaker, and Shalabh (2016). *Introduction to Statistics and Data Analysis - With Exercises, Solutions and Applications in R*. Heidelberg: Springer.

Hjort, L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*, 879–945.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: a tutorial. *Statistical Science 14*, 382–417.

Liboschik, T., K. Fokianos, and R. Fried (2017). tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software 82*(5), 1–51.

Luque Fernandez, M. A., M. Schomaker, B. Rachet, and M. E. Schnitzer (2018). Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine in press*.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. Chapman and Hall/CRC.

Milborrow, S. (2017). *earth: Multivariate Adaptive Regression Splines*. R package version 4.6.0.

Ridgeway, G. (2017). *gbm: Generalized Boosted Regression Models*. R package version 2.1.3.

Schomaker, M. (2012). Shrinkage averaging estimation. *Statistical Papers 53*(4), 1015–1034.

Schomaker, M. (2017). *MAMI: Model Averaging (and Model Selection) after Multiple Imputation*. R package version 0.9.10.

Schuler, M. S. and S. Rose (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *Amreican Journal of Epidemiology 185*(1), 65–73.

Tibsharani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 267–288.

Van der Laan, M., E. Polley, and A. Hubbard (2008). Super learner. *Statistical Applications in Genetics and Molecular Biology 6*, Article 25.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R* (2 ed.). Chapman and Hall/CRC.