

The short-term association between environmental variables and mortality: evidence from Europe

Insights from fine-grained mortality data

Jens Robben Katrien Antonio Torsten Kleinow

September 10, 2024



Motivation and a few basics

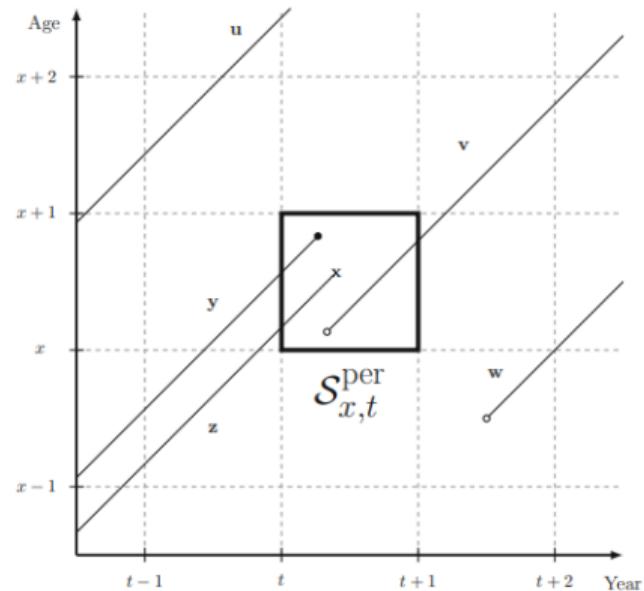
Consider the valuation of life contingent risks, e.g., a life annuity or a whole life insurance product.

Need to master lifetime (or survival) distributions when valuing life-contingent risks.

$q_{x,t,g}$ is called the mortality rate for an x year old in period t , gender g :

$$q_{x,t,g} = 1 - p_{x,t,g} = 1 - \exp(-\mu_{x,t,g}),$$

where $\mu_{x,t,g}$ is the force of mortality and assumed piecewise constant.



- ▶ Epidemiological studies have unveiled (short-term) associations between **mortality** and
 - **temperature**, e.g., Keatinge et al. [2000] and Basu and Samet [2002],
 - **cold spells and heat waves**, e.g., Braga et al. [2001] and Pattenden et al. [2003],
 - **air pollution**, e.g., Pascal et al. [2014] for PM10 and PM2.5 and Orellano et al. [2020] for ozone and nitrogen dioxide.
- ▶ Various methodologies have been proposed:
 - **Poisson regression models**, e.g., Armstrong [2006] and Braga et al. [2002],
 - **Distributed Lag (Non-Linear) Models**, e.g., Schwartz [2000] and Gasparrini et al. [2010],
 - **Extreme value analysis**, e.g., Li and Tang [2022].

I will present today our working paper:

The association between environmental variables and short-term mortality: evidence from Europe

working paper on [arxiv](#), code on [GitHub](#), by Jens Robben (UvA, RCLR), Katrien Antonio and Torsten Kleinow (UvA, RCLR).

1. Identify the primary environmental factors contributing to the estimation of mortality deviations from a pre-defined baseline level.
2. Investigate the marginal impact of an environmental factor on deviations from the mortality baseline level.
3. Investigate how environmental factors interact when modelling mortality rates. Are there harvesting effects present?
4. Quantify the added contribution of environmental factors when aggregating the resulting estimated weekly mortality rates on an annual or country-level basis.

Short-term association between environmental variables and mortality rates

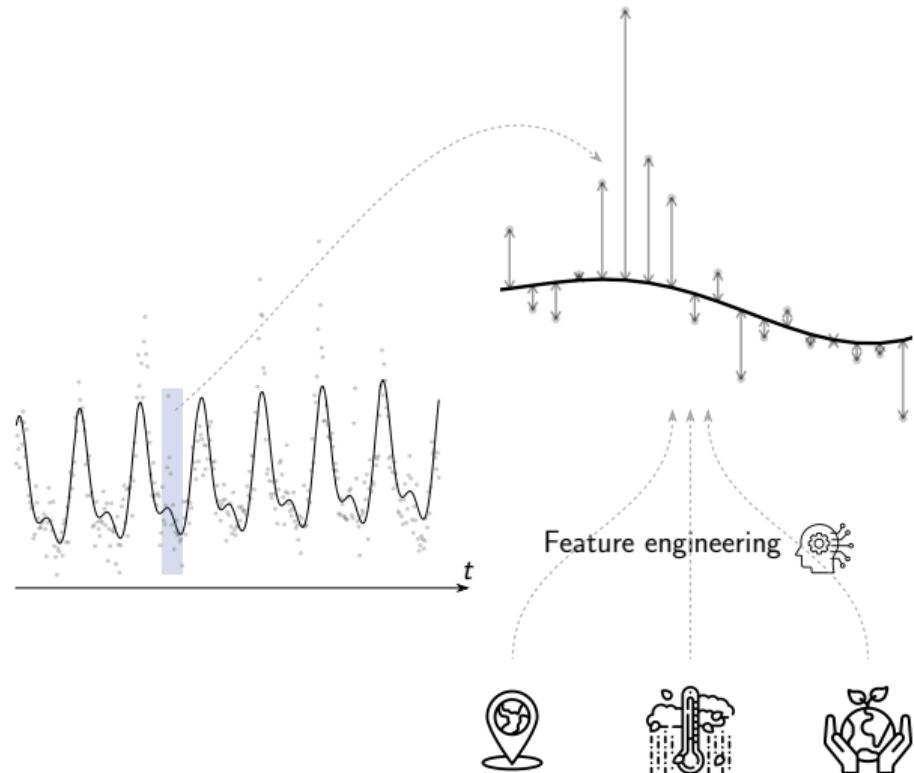
Introduction

Plan of attack

Using fine-grained **open data**, study the association between **environmental factors** and **weekly mortality rates** in European regions.

Proposed framework:

1. a weekly, region-specific **baseline** mortality model to capture overall **seasonal** trends.
2. a predictive model to analyze **mortality deviations** from the baseline model using region-specific environmental factors.



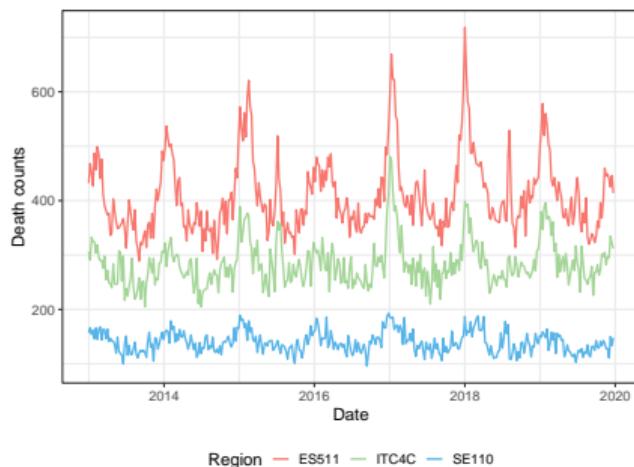
Data

Death counts

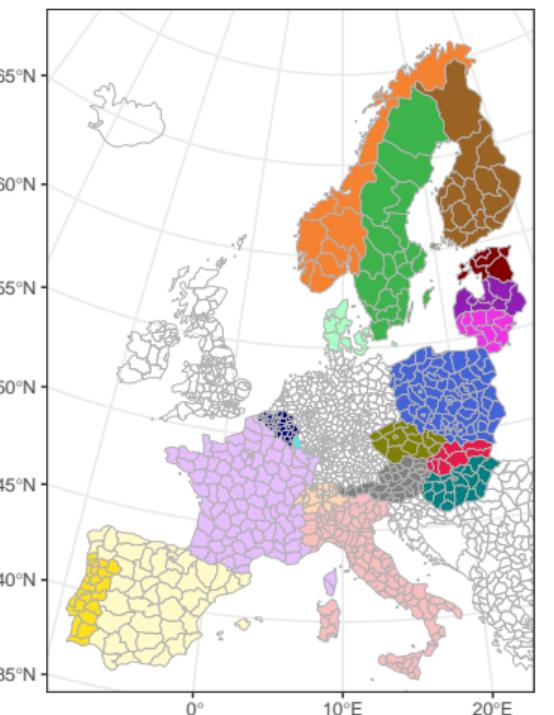
Eurostat: deaths by week, gender, 5-year age group and NUTS 3 region from 20 European countries throughout the years 2013-2019 (> 500 regions).

Focus on old age group 65+, unisex.

Seasonal trend:



NUTS 3 regions



AT	BE	CH	CZ	DK	EE	ES
FI	FR	HU	IT	LI	LT	LU
LV	NO	PL	PT	SE	SK	

Data

Weather data

E-OBS land-only, gridded meteorological data for Europe from the Copernicus Climate Data Store.

Daily, high-resolution gridded dataset, defined on a grid with a spatial resolution of 0.10° (≈ 9 km).

Weather factors:

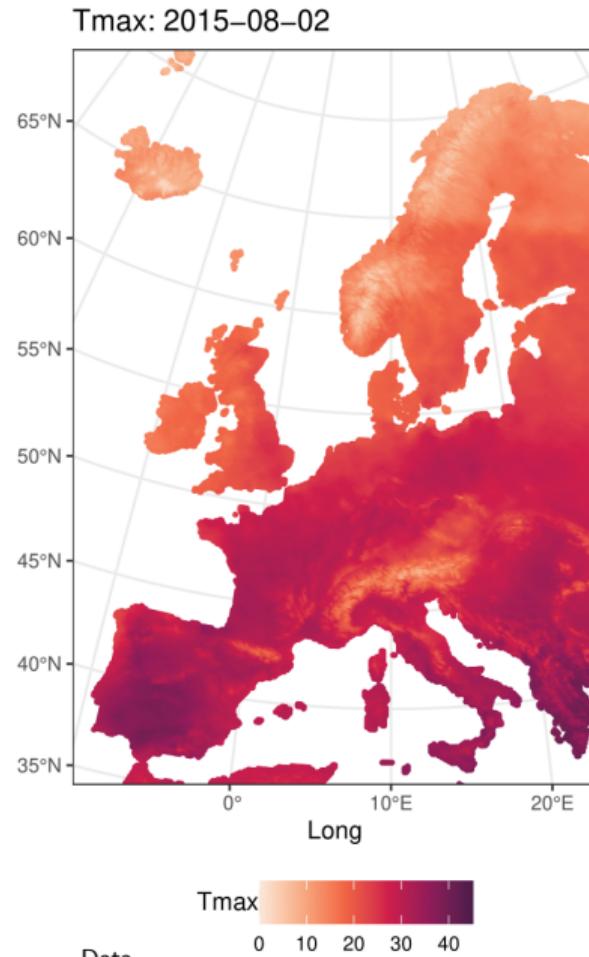
Tmax: daily maximum temperature

Tmin: daily minimum temperature

Hum: daily average relative humidity

Rain: total daily precipitation

Wind: daily average wind speed



Weather data

E-OBS land-only, gridded meteorological data for Europe from the Copernicus Climate Data Store.

Daily, high-resolution gridded dataset, defined on a grid with a spatial resolution of 0.10° (≈ 9 km).

Weather factors:

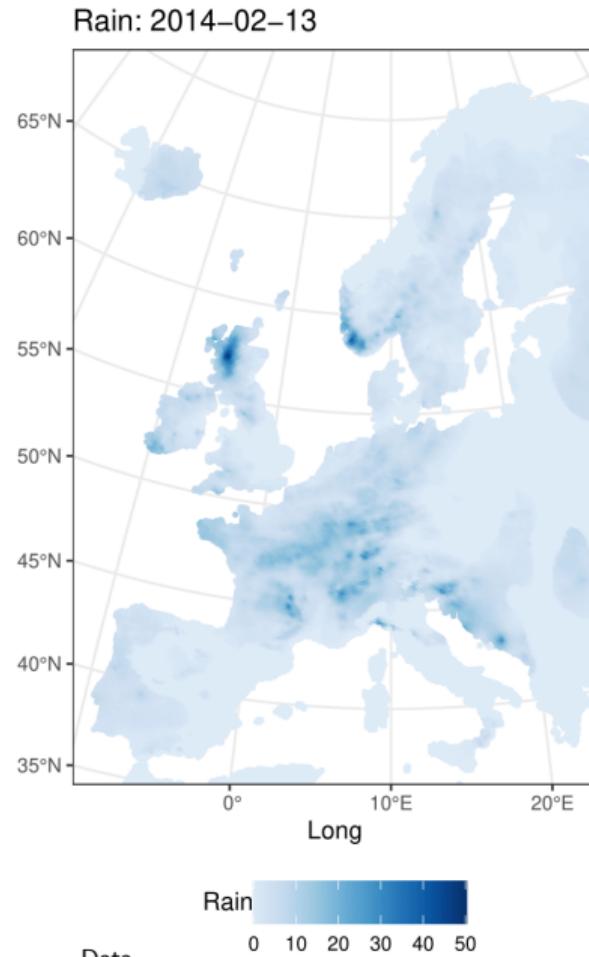
Tmax: daily maximum temperature

Tmin: daily minimum temperature

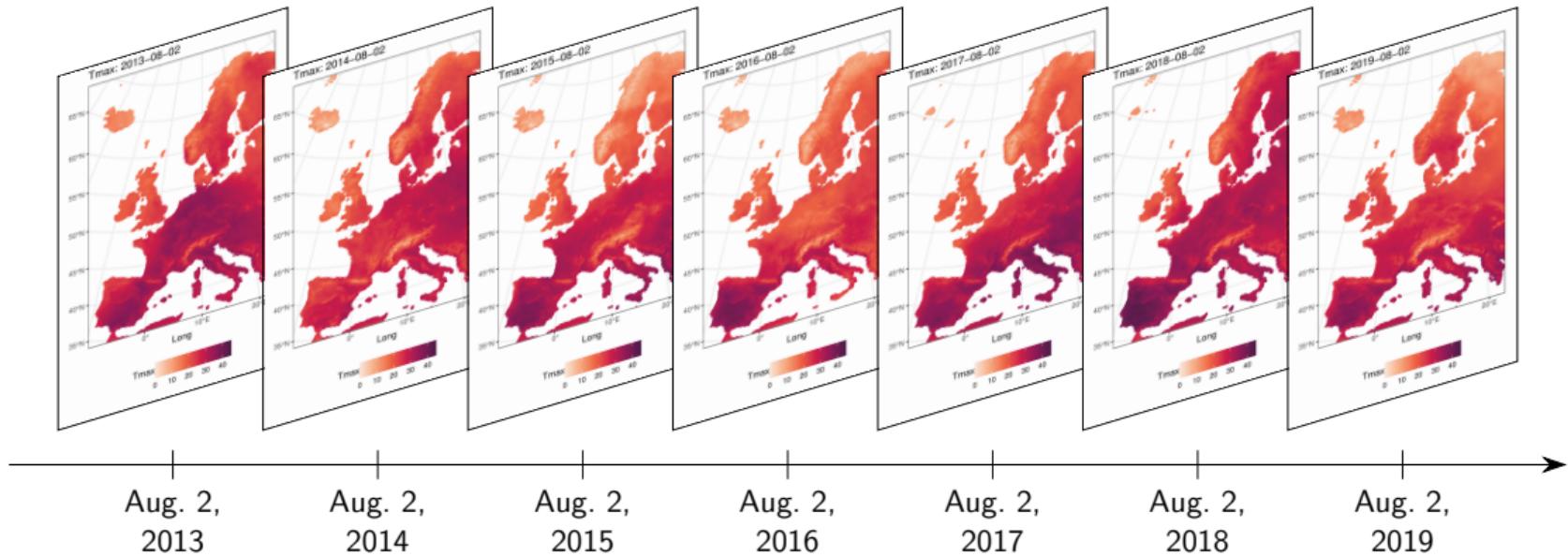
Hum: daily average relative humidity

Rain: total daily precipitation

Wind: daily average wind speed



Weather data



Air pollution data

CAMS European air quality reanalyses dataset from the Copernicus Atmosphere Monitoring Service (land + sea).

Hourly, high-resolution air quality reanalyses, defined on a grid with a spatial resolution of 0.10° (≈ 9 km).

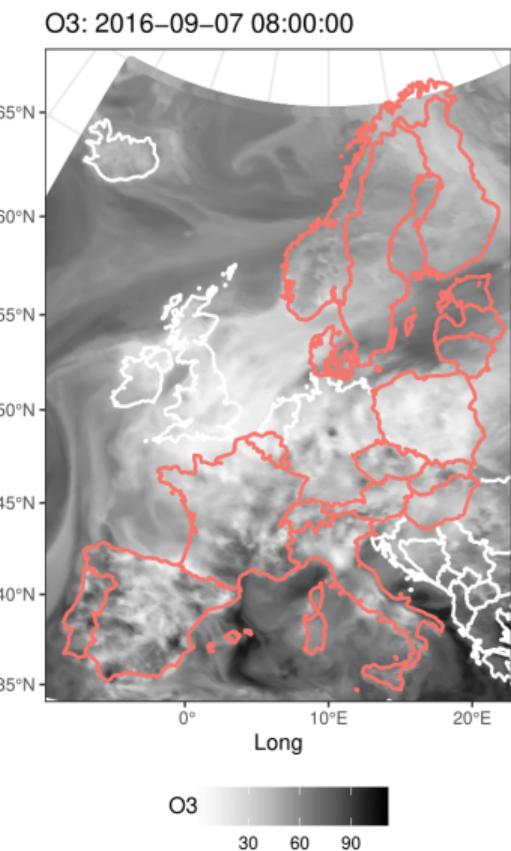
Air pollutants ($\mu\text{g}/\text{m}^3$):

O3: hourly ozone levels.

NO₂: hourly nitrogen dioxide levels.

PM10: hourly particular matter (10 microns wide).

PM2.5: hourly particular matter (2.5 microns wide).



Air pollution data

CAMS European air quality reanalyses dataset from the Copernicus Atmosphere Monitoring Service (land + sea).

Hourly, high-resolution air quality reanalyses, defined on a grid with a spatial resolution of 0.10° (≈ 9 km).

Air pollutants ($\mu\text{g}/\text{m}^3$):

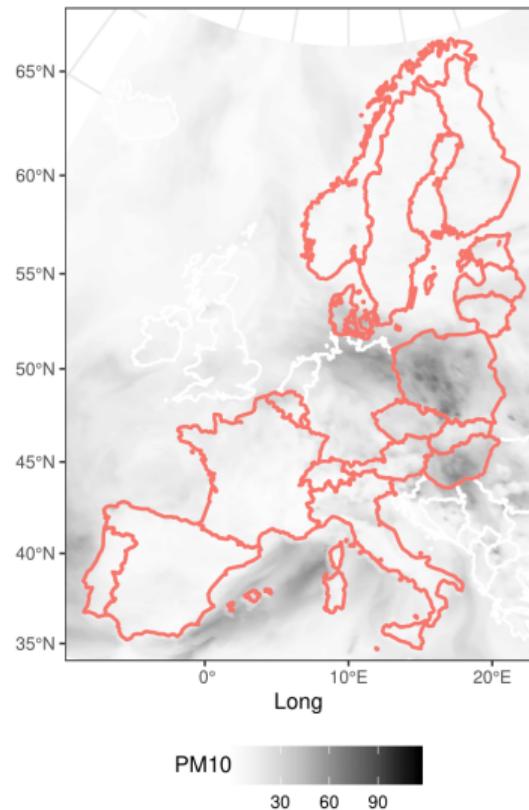
O3: hourly ozone levels.

NO₂: hourly nitrogen dioxide levels.

PM10: hourly particular matter (10 microns wide).

PM2.5: hourly particular matter (2.5 microns wide).

PM10: 2019-02-01 10:00:00



Model specification

Weekly, region-specific baseline mortality model

A weekly, region-specific baseline mortality model to capture the overall seasonal trends in the considered regions.

Incorporate seasonality through Fourier terms

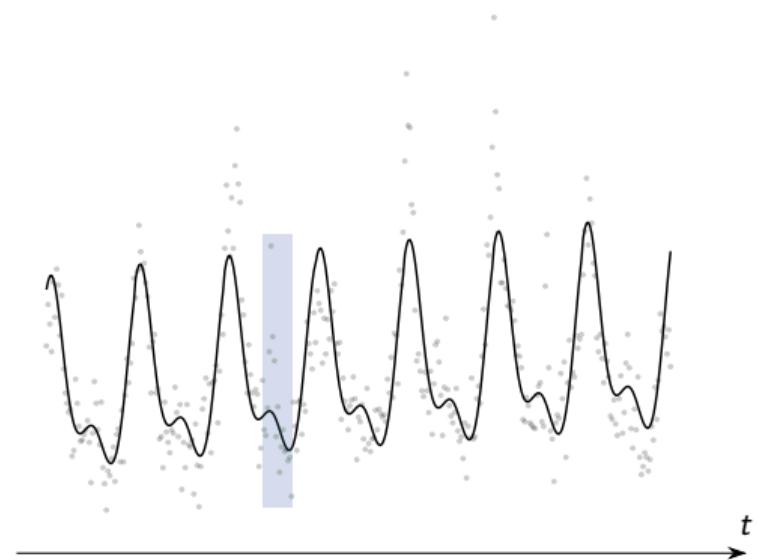
Serfling [1963]:

$$D_{t,w}^{(r)} \sim \text{Poisson} \left(E_{t,w}^{(r)} \cdot \mu_{t,w}^{(r)} \right),$$

$$\log \mu_{t,w}^{(r)} = \beta_0^{(r)} + \beta_1^{(r)} t + \beta_2^{(r)} \sin \left(\frac{2\pi w}{52} \right) + \beta_3^{(r)} \cos \left(\frac{2\pi w}{52} \right) + \\ \beta_4^{(r)} \sin \left(\frac{2\pi w}{26} \right) + \beta_5^{(r)} \cos \left(\frac{2\pi w}{26} \right).$$

Region-specific population exposures $E_{t,w}^{(r)}$ from Eurostat.

Estimated baseline death counts: $\hat{b}_{t,w}^{(r)} := E_{t,w}^{(r)} \cdot \hat{\mu}_{t,w}^{(r)}$.



Modelling mortality deviations from the baseline model

Explain observed deviations from the baseline deaths using region-specific environmental features.

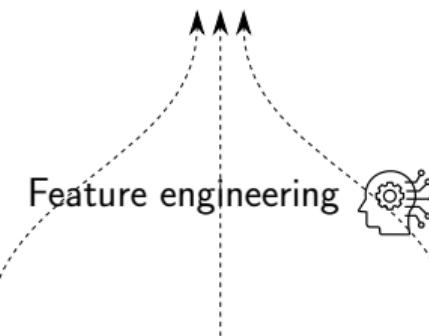
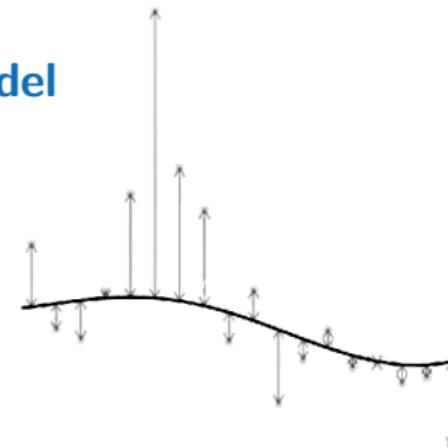
Fix estimated baseline deaths and impose distributional assumption:

$$D_{t,w}^{(r)} \sim \text{Poisson} \left(\hat{b}_{t,w}^{(r)} \phi_{t,w}^{(r)} \right),$$

$$\phi_{t,w}^{(r)} = f \left(\text{long}^{(r)}, \text{lat}^{(r)}, \text{season}_{t,w}, \mathbf{c}_{t,w}^{(r)}, \mathbf{e}_{t,w}^{(r)}, I^1 \left(\mathbf{c}_{t,w}^{(r)} \right), I^1 \left(\mathbf{e}_{t,w}^{(r)} \right), \dots, I^s \left(\mathbf{c}_{t,w}^{(r)} \right), I^s \left(\mathbf{e}_{t,w}^{(r)} \right) \right).$$

$f(\cdot)$ is a selected predictive modelling technique.

We opt for a machine learning model.



Model specification

Model calibration

Calibrating the baseline model

Fit a **Poisson GLM** jointly across all regions, add a smoothness penalty:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(-\log \mathcal{L}_P(\beta) + \sum_{p=0}^5 \lambda_p \beta_p^T \mathbf{S} \beta_p \right),$$

where:

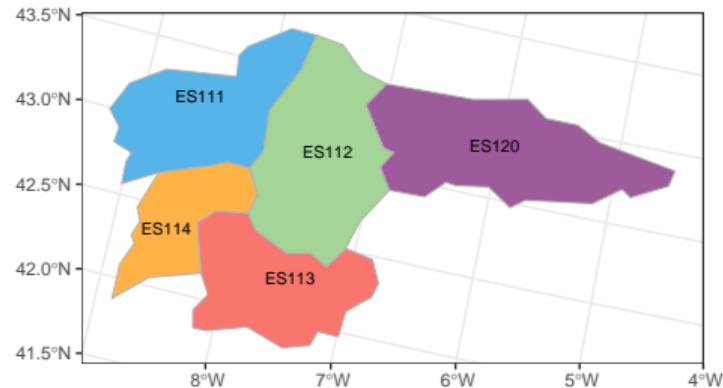
β : parameter vector

$\log \mathcal{L}_P(\beta)$: Poisson log-likelihood

$\beta_p^T \mathbf{S} \beta_p$: penalizes sum of squared differences between coefs of adjacent regions

λ_p : smoothing or penalty parameter.

Example (5 Spanish NUTS 3 regions):



Penalty matrix \mathbf{S} :

	ES111	ES112	ES113	ES114	ES120
ES111	2	-1	0	-1	0
ES112	-1	4	-1	-1	-1
ES113	0	-1	2	-1	0
ES114	-1	-1	-1	3	0
ES120	0	-1	0	0	1

Calibrating the mortality deviations model

XGBoost: flexible and efficient implementation of gradient boosting.

Tuning parameters:

`nrounds`: number of boosting iterations.

`eta`: learning rate.

`max_depth`: the maximum depth of a tree.

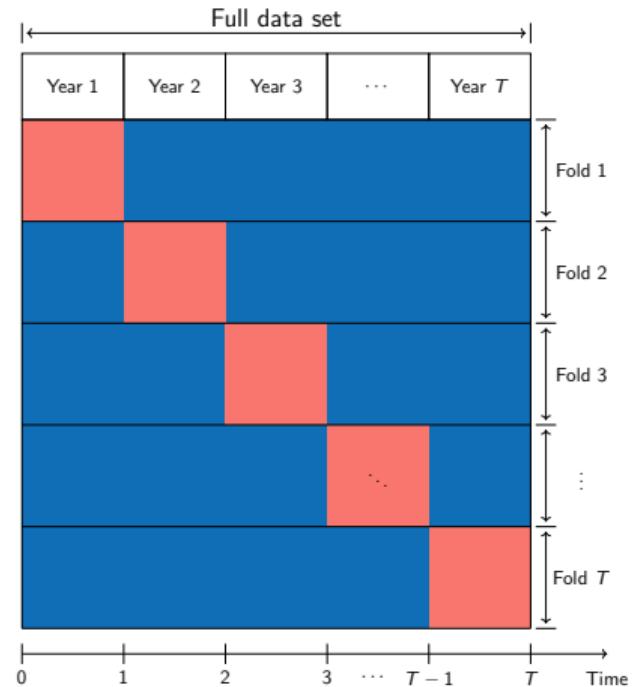
`subsample`: subsample ratio of the training data.

`colsample_bytree`: subsample ratio of the features.

Parameter tuning with T-fold cross-validation.

Calibrate XGBoost model on entire training data with optimal parameter configuration.

Interpretation tools to gain insights: VIP, ALE.



Feature engineering

Feature engineering

Motivation

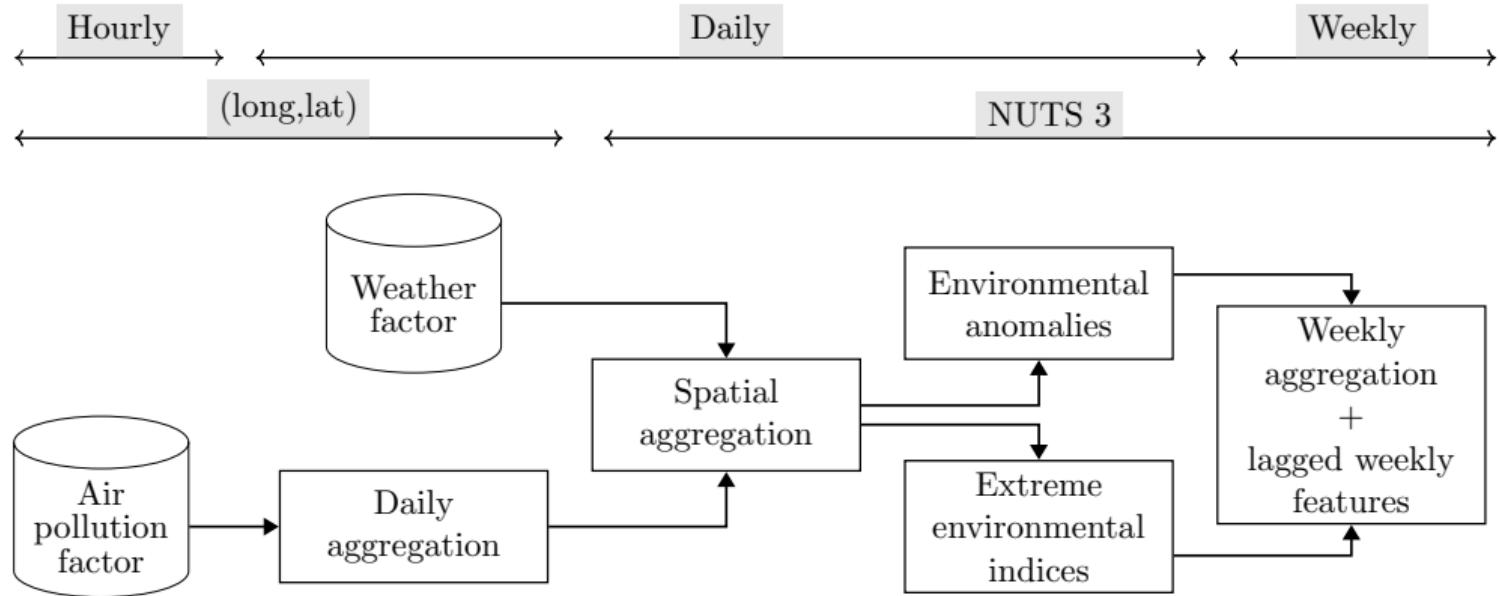
Differences in spatial and temporal dimension across the data sources:

- data on death counts: weekly, NUTS 3 scale.
- environmental data: hourly or daily time scale, spatial grid.

Goal of feature engineering:

- convert the temporal and spatial dimensions of the environmental data into features on a weekly, NUTS 3 scale
- create features that measure deviations from baseline conditions from environmental data to explain excess or deficit mortality (other ideas welcome!).

Flow chart



Aim: to capture the effects of extreme environmental conditions on mortality baseline deviations.

Calculate region-specific 5% and 95% quantiles of the daily historical temperature or air pollution observations over the years 2013-2019.

Define extreme high temperature index (hot-day index):

$$T.ind_{t,w,d}^{(r,95\%)} = \mathbb{1} \left\{ T_{\max,t,w,d}^{(r)} \geq q_{T_{\max}}^{(r,95\%)} \right\} + \mathbb{1} \left\{ T_{\text{avg},t,w,d}^{(r)} \geq q_{T_{\text{avg}}}^{(r,95\%)} \right\} + \mathbb{1} \left\{ T_{\min,t,w,d}^{(r)} \geq q_{T_{\min}}^{(r,95\%)} \right\}.$$

Index values: 0-3, indicating the severity of hot days.

Similar extreme indices are created for the other daily weather and air pollution factors.

Environmental anomalies

Create features that **quantify deviations from typical, baseline conditions** for each day throughout the year.

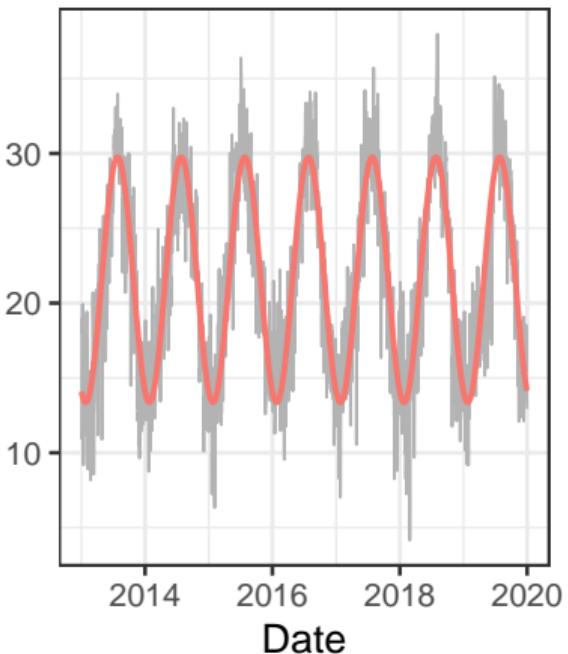
Robust linear regression to capture baseline:

$$\tilde{x}_{t,w,d}^{(r)} = \alpha_0^{(r)} + \alpha_1^{(r)} \sin\left(\frac{2\pi d}{365.25}\right) + \alpha_2^{(r)} \cos\left(\frac{2\pi d}{365.25}\right) + \epsilon_{t,w,d}^{(r)},$$

In the paper, we work with excesses or deviations from the baseline (**anomalies**):

$$\tilde{x}_{t,w,d}^{(r)} - \hat{x}_{t,w,d}^{(r)}$$

ES511: Barcelona



Environmental anomalies

Create features that [quantify deviations from typical, baseline conditions](#) for each day throughout the year.

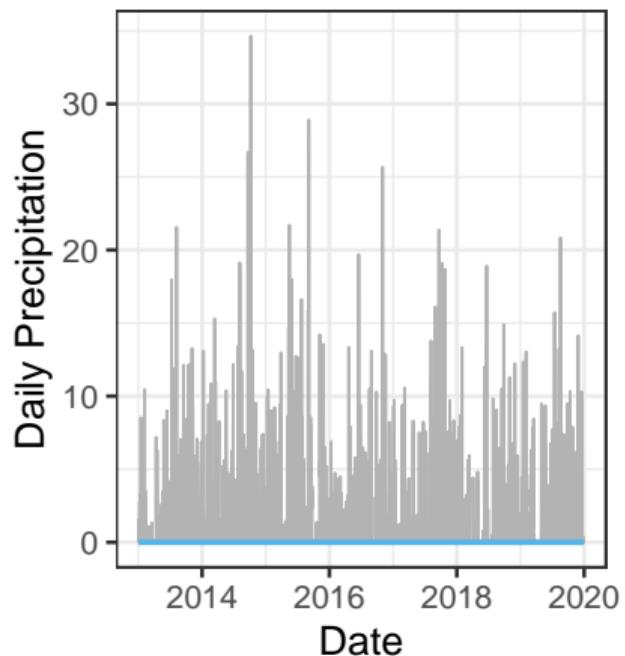
[Robust linear regression](#) to capture baseline:

$$\tilde{x}_{t,w,d}^{(r)} = \alpha_0^{(r)} + \alpha_1^{(r)} \sin\left(\frac{2\pi w}{365.25}\right) + \alpha_2^{(r)} \cos\left(\frac{2\pi w}{365.25}\right) + \epsilon_{t,w,d}^{(r)},$$

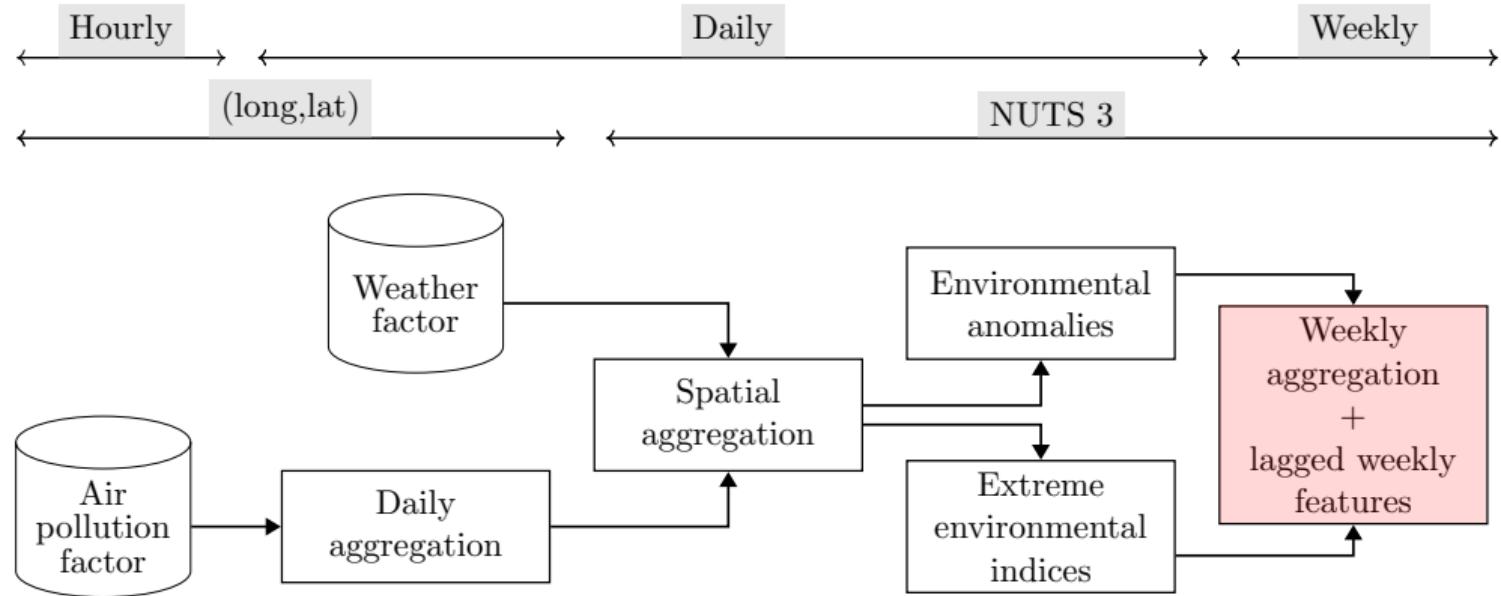
In the paper we work with excesses or deviations from the baseline ([anomalies](#)):

$$\tilde{x}_{t,w,d}^{(r)} - \hat{x}_{t,w,d}^{(r)}$$

SE110: Stockholms län

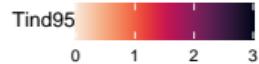
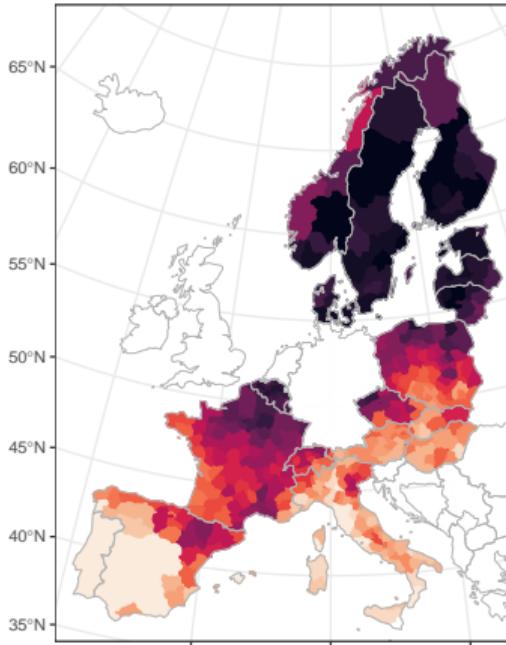


Flow chart

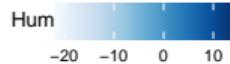
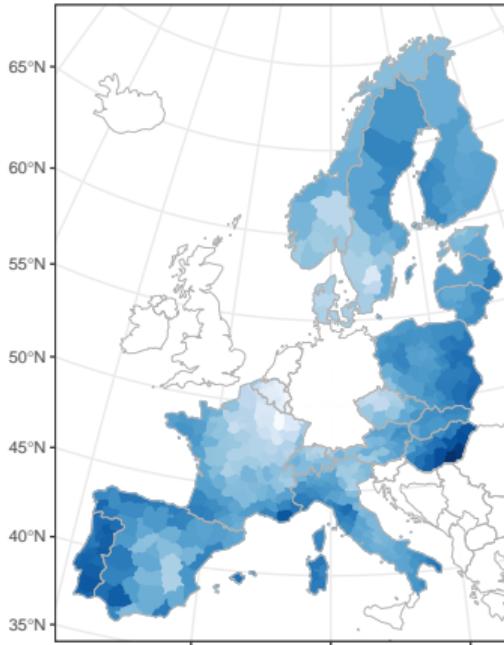


Weekly aggregation

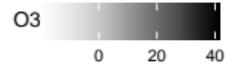
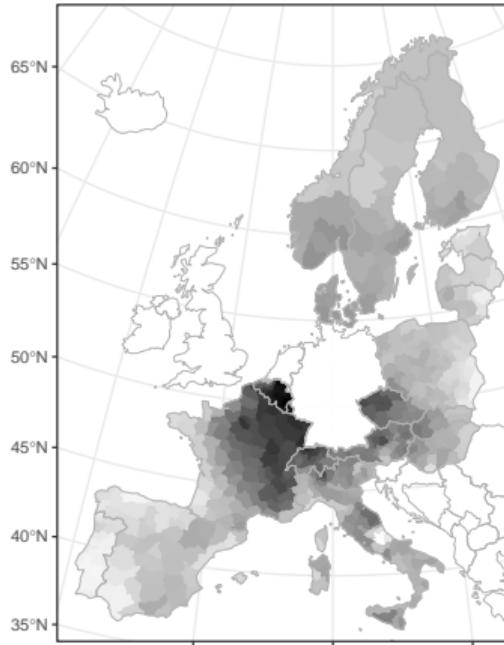
w_avg_Tind95: 2018–30



w_avg_Hum_anom: 2018–30

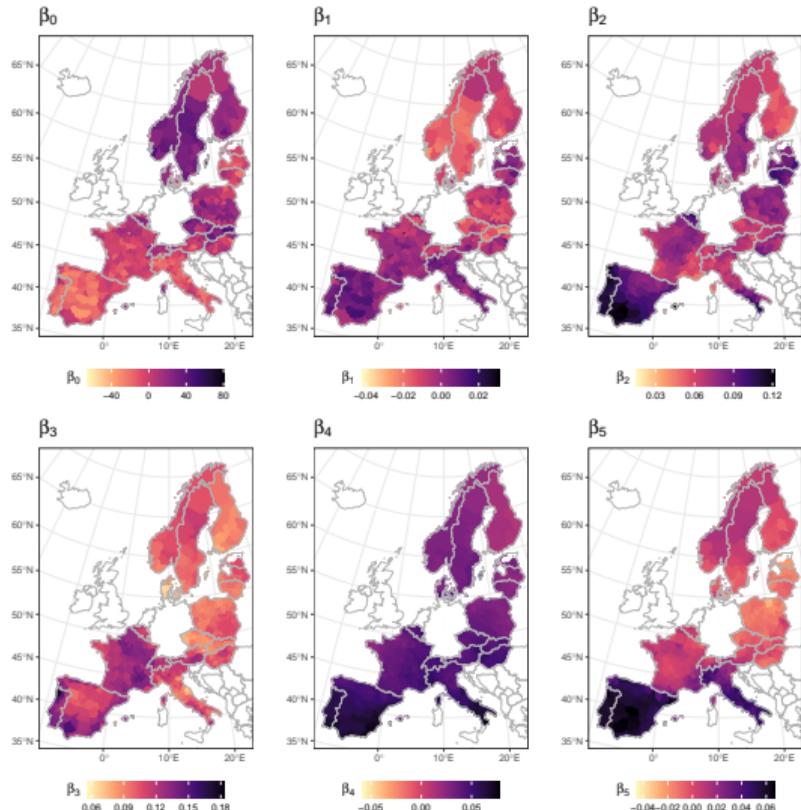


w_avg_O3_anom: 2018–30



Calibration results

Baseline model



Machine learning model

Input features: longitude-latitude coordinates, season, (one-week lagged) environmental anomalies and extreme indices.

Tuning by 7-fold cross validation over the years 2013-2019 using an extensive tuning grid.

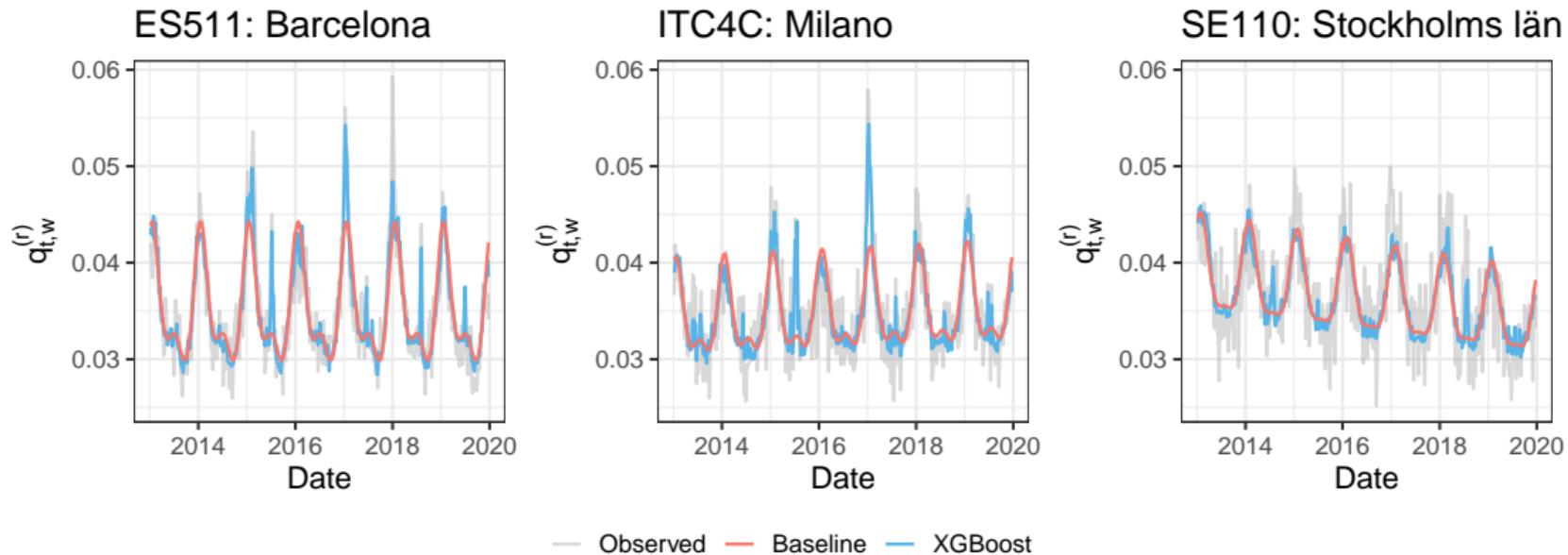
Tuning parameters: `nrounds` (490), `eta` (0.01), `min_child_weight` (1000), `max_depth` (7), `subsample` (0.75), `colsample_bytree` (0.50).

Insights in the machine-learning model

In-sample fit and model performance

21

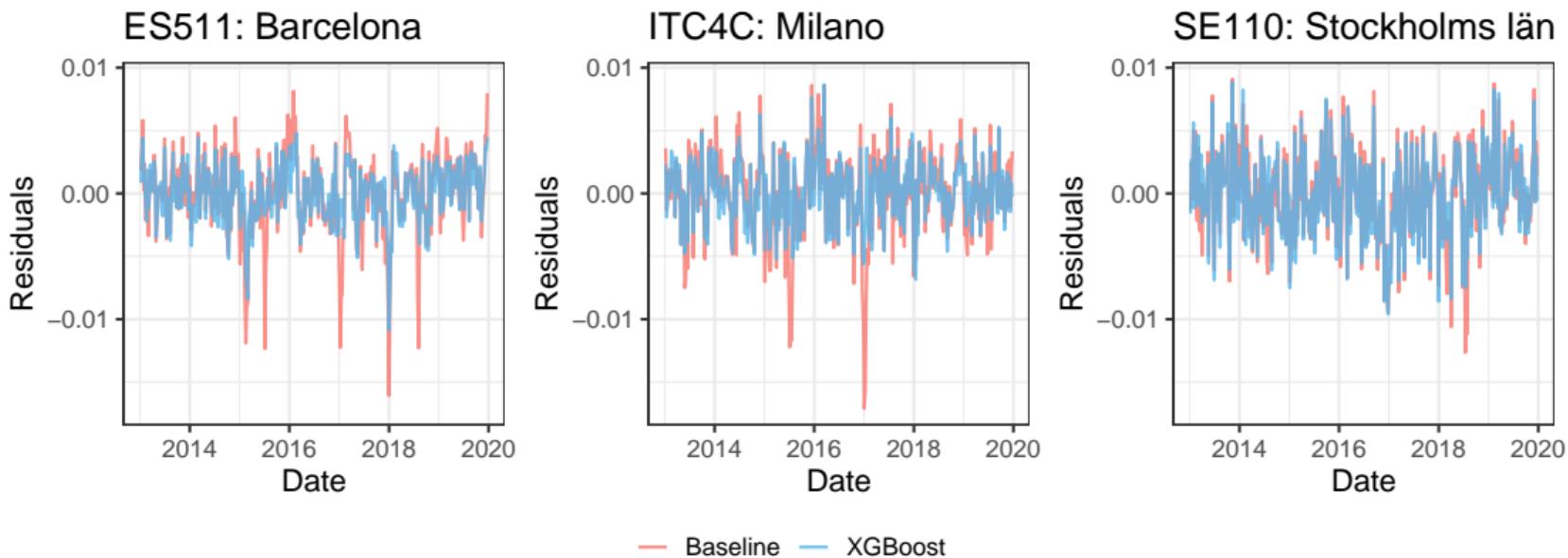
Observed and estimated mortality rates (baseline + XGBoost):



In-sample fit and model performance

22

Residuals of the estimated weekly mortality rates (baseline + XGBoost):



Machine learning techniques perform **automatic feature selection**.

Which features do significantly contribute to the predictions?

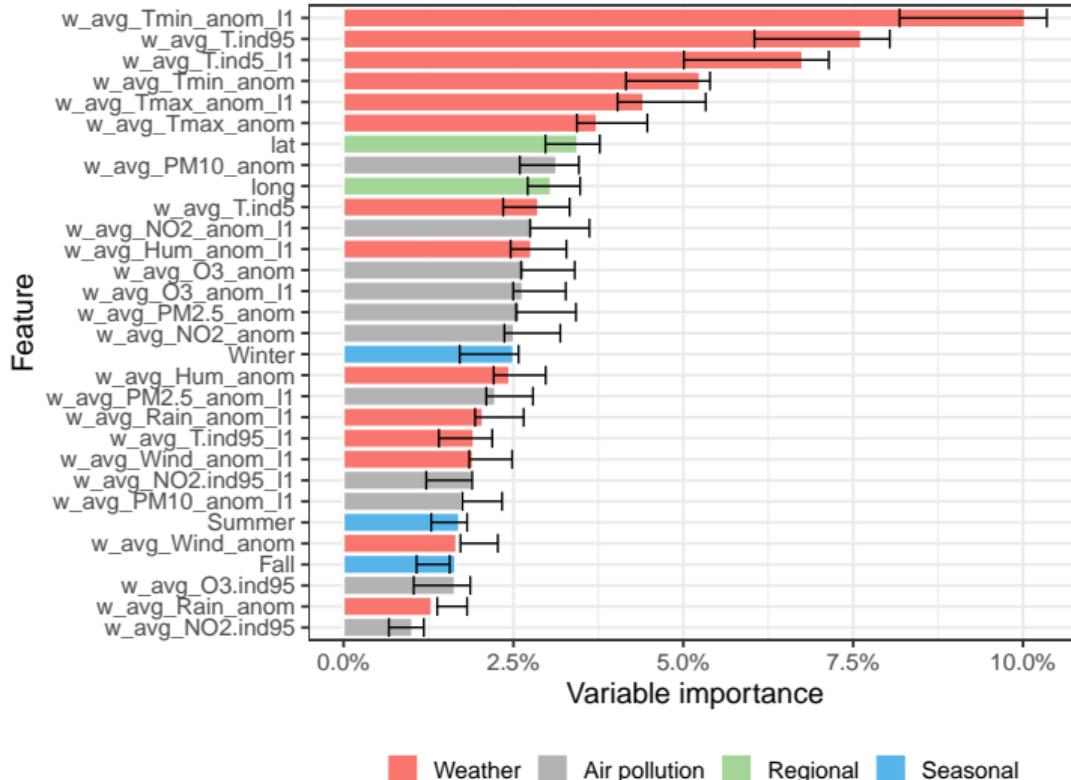
We calculate the **feature importance** of each feature X_I as:

$$V_{\text{imp}}(X_I) = \frac{1}{\text{nrounds}} \sum_{n=1}^{\text{nrounds}} \Delta \mathcal{L}_n(X_I),$$

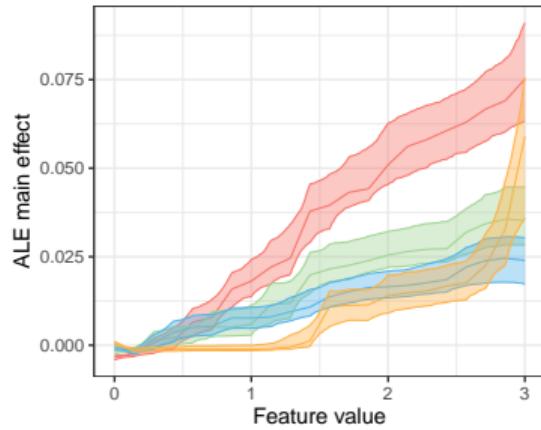
with $\Delta \mathcal{L}_n(X_I)$ the total reduction in the Poisson loss function, caused by splits associated to feature X_I in the tree built during iteration n of the XGBoost algorithm.

Features with a high importance appear **often** and **high** in the tree.

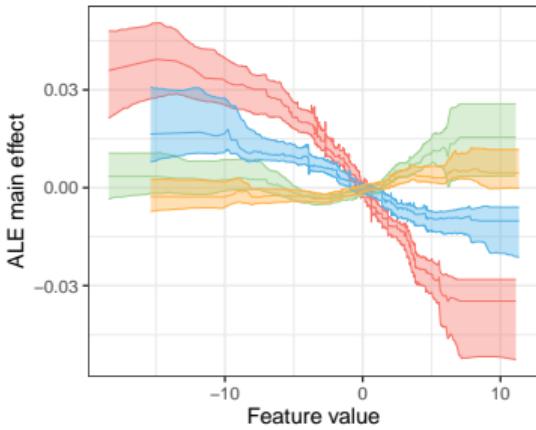
Feature importance



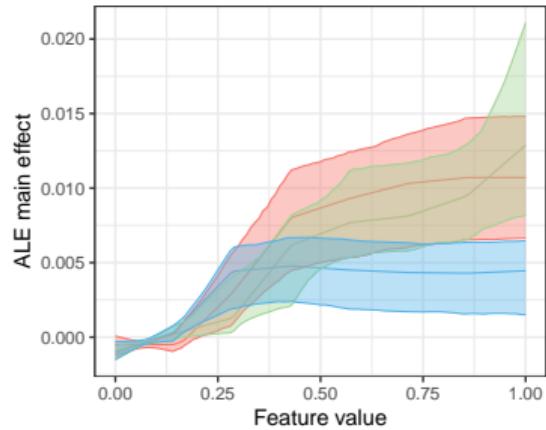
ALE main effects



■ T.ind95 (7.62%) □ T.ind5 (2.86%)
■ T.ind5_I1 (6.76%) □ T.ind95_I1 (1.92%)



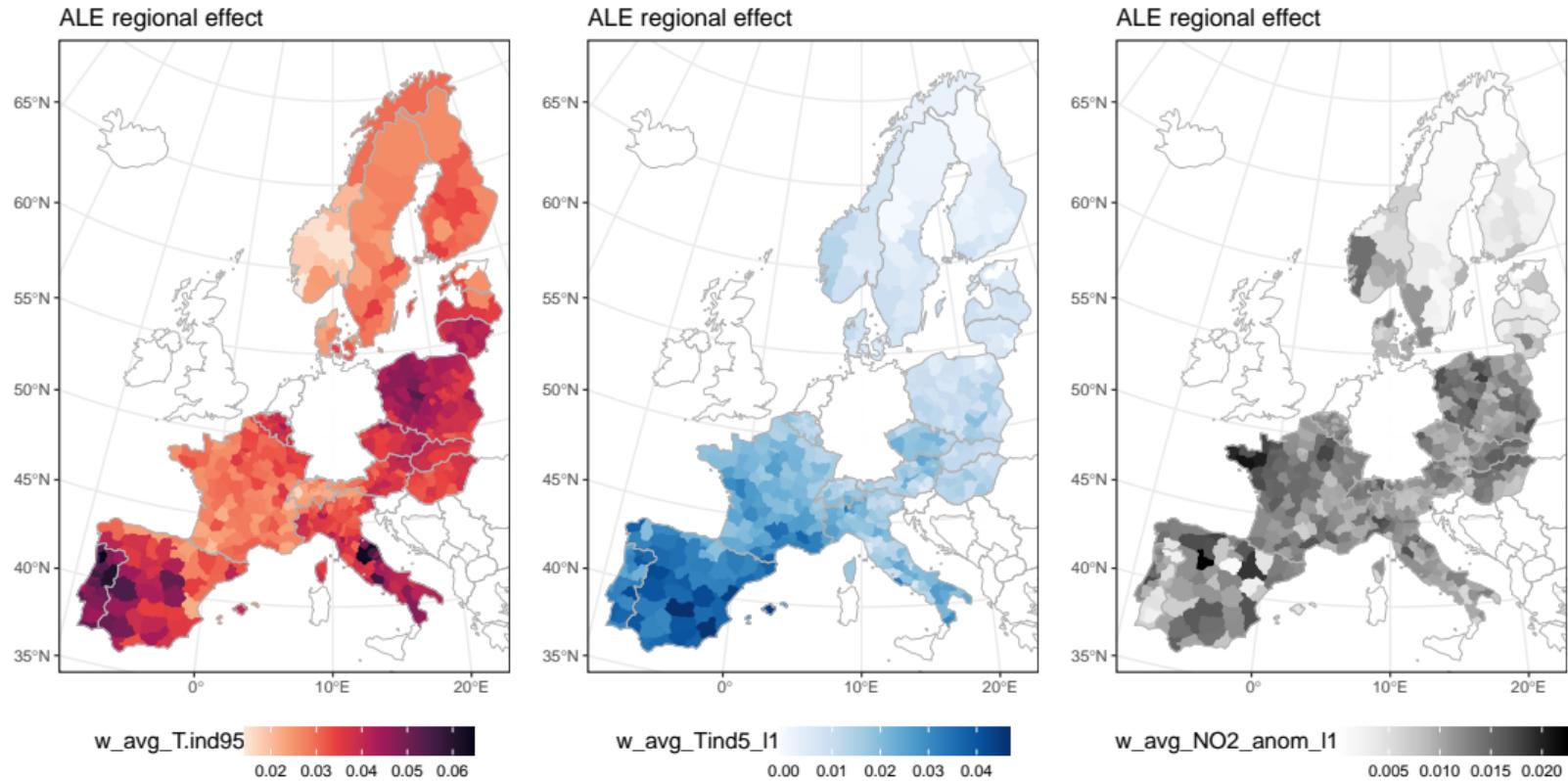
■ Tmin_anom_I1 (10.03%) □ Tmax_anom_I1 (4.41%)
■ Tmin_anom (5.24%) □ Tmax_anom (3.73%)



■ NO2.ind95_I1 (1.88%) □ NO2.ind95 (1.01%)
■ O3.ind95 (1.64%)

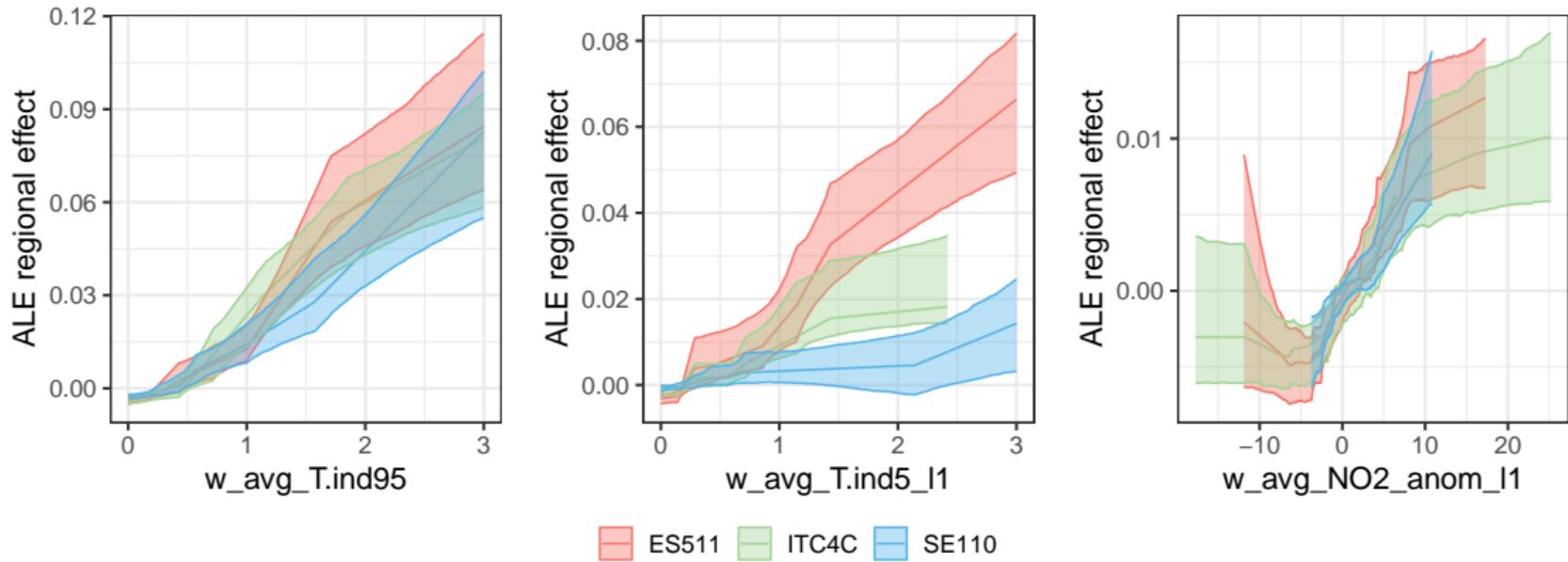
ALE regional effects

26

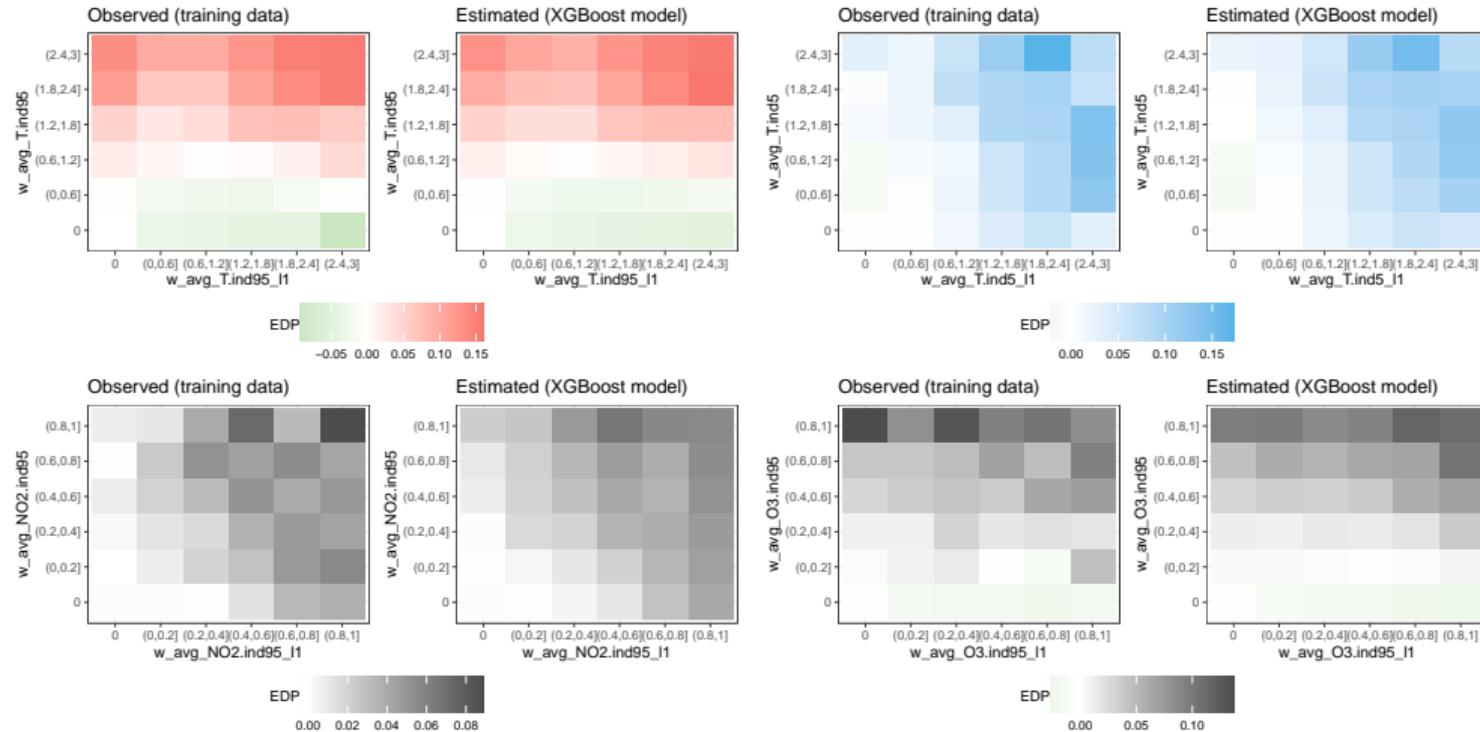


ALE regional effects

27



Harvesting effects



Wrap up

We have (multiple) additional visuals and analyses in the **working paper**, as well as a detailed discussion of related literature.

Limitation: focus on short-term associations only!

Exciting opportunities ahead using sophisticated learning methods and **fine-grained (open) data**.

However, feature engineering, (proper) interpretation tools, **interplay** of more traditional actuarial + statistical learning and sophisticated machine learning methods are key!

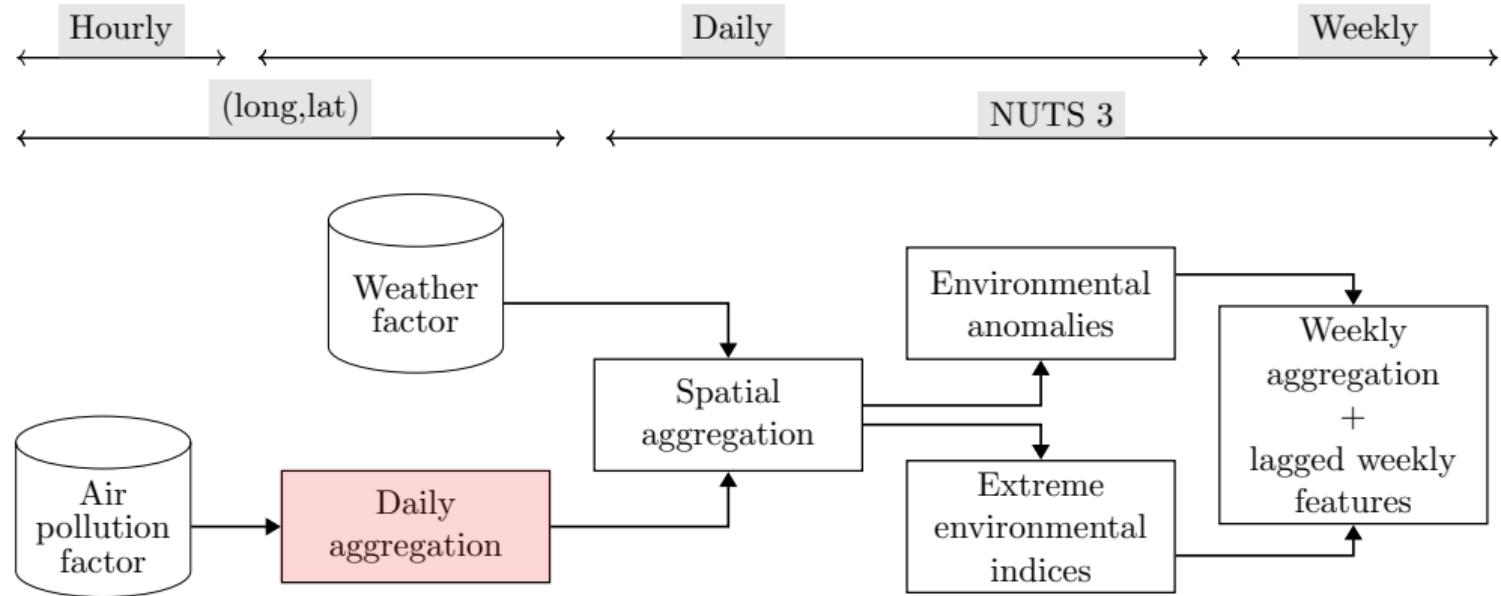
- Ben Armstrong. Models for the relationship between ambient temperature and daily mortality. *Epidemiology*, pages 624–631, 2006.
- Rupa Basu and Jonathan M Samet. Relation between elevated ambient temperature and mortality: a review of the epidemiologic evidence. *Epidemiologic reviews*, 24(2):190–202, 2002.
- Alfésio LF Braga, Antonella Zanobetti, and Joel Schwartz. The effect of weather on respiratory and cardiovascular deaths in 12 us cities. *Environmental health perspectives*, 110(9):859–863, 2002.
- Alfésio Luís Ferreira Braga, Antonella Zanobetti, and Joel Schwartz. The time course of weather-related deaths. *Epidemiology*, 12(6):662–667, 2001.
- Antonio Gasparrini, Ben Armstrong, and Mike G Kenward. Distributed lag non-linear models. *Statistics in medicine*, 29(21):2224–2234, 2010.

- William R Keatinge, Gavin C Donaldson, Elvira Cordioli, Martina Martinelli, Anton E Kunst, Johan P Mackenbach, Simo Nayha, and Ilkka Vuori. Heat related mortality in warm and cold regions of europe: observational study. *Bmj*, 321(7262):670–673, 2000.
- Han Li and Qihe Tang. Joint extremes in temperature and mortality: A bivariate pot approach. *North American Actuarial Journal*, 26(1):43–63, 2022.
- Pablo Orellano, Julieta Reynoso, Nancy Quaranta, Ariel Bardach, and Agustin Ciapponi. Short-term exposure to particulate matter (pm10 and pm2. 5), nitrogen dioxide (no2), and ozone (o3) and all-cause and cause-specific mortality: Systematic review and meta-analysis. *Environment international*, 142:105876, 2020. doi: 10.1016/j.envint.2020.105876.
- Mathilde Pascal, Grégoire Falq, Vérène Wagner, Edouard Chatignoux, Magali Corso, Myriam Blanchard, Sabine Host, Laurence Pascal, and Sophie Larrieu. Short-term impacts of particulate matter (pm10, pm10–2.5, pm2. 5) on mortality in nine french cities. *Atmospheric Environment*, 95:175–184, 2014. doi: 10.1016/j.atmosenv.2014.06.030.

References III

- S Pattenden, B Nikiforov, and B Armstrong. Mortality and temperature in sofia and london. *Journal of Epidemiology and Community health*, 57(8):628, 2003.
- Joel Schwartz. The distributed lag between air pollution and daily deaths. *Epidemiology*, 11(3):320–326, 2000.
- Robert E Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public health reports*, 78(6):494, 1963.

Flow chart



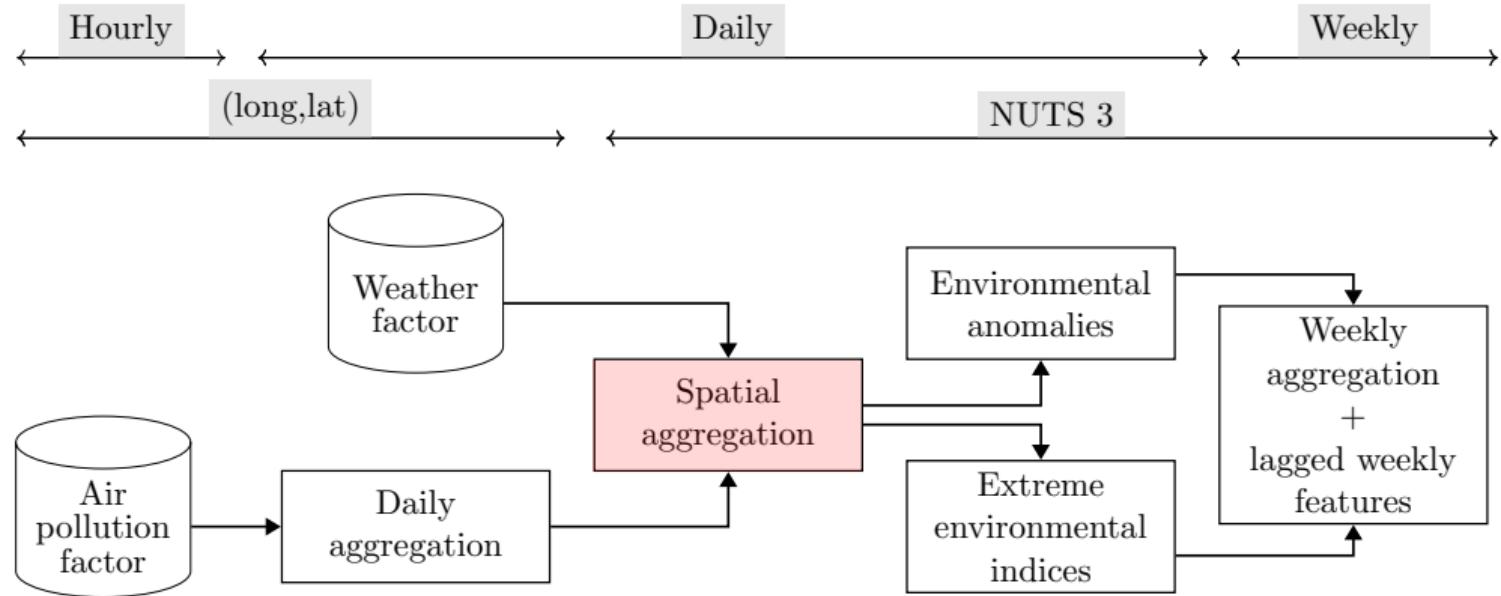
Consider an air pollution factor and denote its concentration at hour h of day d in week w of year t and located at longitude-latitude coordinates (long,lat) as $x_{t,w,d,h}^{(\text{long},\text{lat})}$.

Compute the **daily minimum, average, and maximum concentrations** of the air pollutant, measured at the coordinates (long,lat) as:

$$\begin{aligned}\hat{x}_{t,w,d}^{(\text{long},\text{lat})} &= \min \left\{ x_{t,w,d,h}^{(\text{long},\text{lat})} \mid h = 0, 1, \dots, 23 \right\} \\ \bar{x}_{t,w,d}^{(\text{long},\text{lat})} &= \text{avg} \left\{ x_{t,w,d,h}^{(\text{long},\text{lat})} \mid h = 0, 1, \dots, 23 \right\} \\ \vee{x}_{t,w,d}^{(\text{long},\text{lat})} &= \max \left\{ x_{t,w,d,h}^{(\text{long},\text{lat})} \mid h = 0, 1, \dots, 23 \right\}.\end{aligned}$$

Weather factors already available at the daily level (no need for daily aggregation).

Flow chart



Spatial aggregation

$\tilde{x}_{t,w,d}^{(\text{long},\text{lat})}$: daily level of a specific environmental feature at coordinates (long, lat) for year t , week w , and day d .

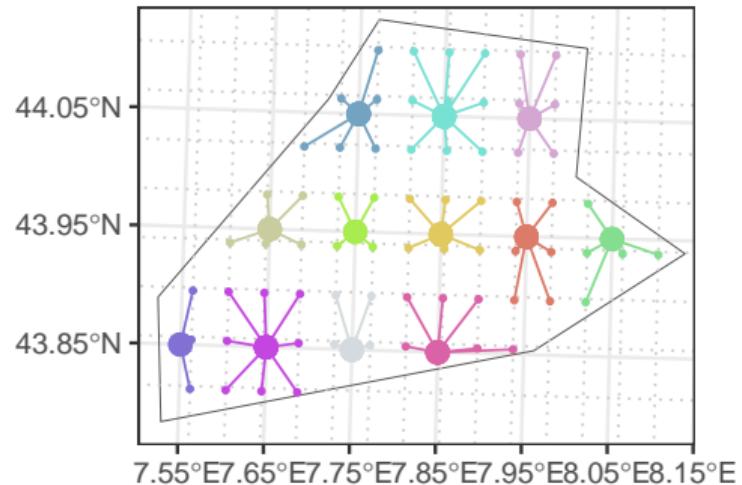
Construct feature on NUTS 3 scale:

$$\tilde{x}_{t,w,d}^{(r)} = \sum_{(\text{long},\text{lat}) \in \mathcal{I}_1(r)} \omega_{(\text{long},\text{lat})} \cdot \tilde{x}_{t,w,d}^{(\text{long},\text{lat})},$$

where:

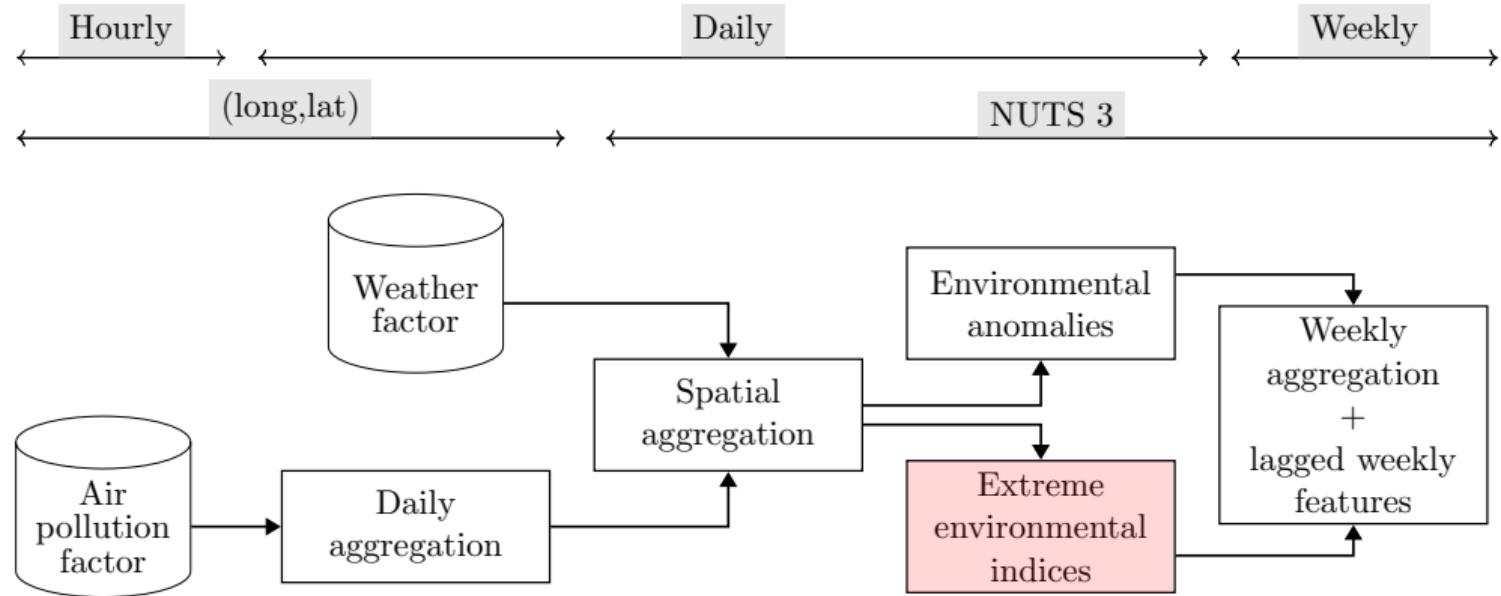
$\omega_{(\text{long},\text{lat})}$: population weights using gridded population data from the Socioeconomic Data and Applications Center (NASA's EOSDIS)

ITC31: Imperia



- Feature grid $I_1(r)$
- Population grid $I_2(r)$

Flow chart



Aim: to capture the effects of extreme environmental conditions on mortality baseline deviations.

Calculate region-specific 5% and 95% quantiles of the daily historical temperature or air pollution observations over the years 2013-2019.

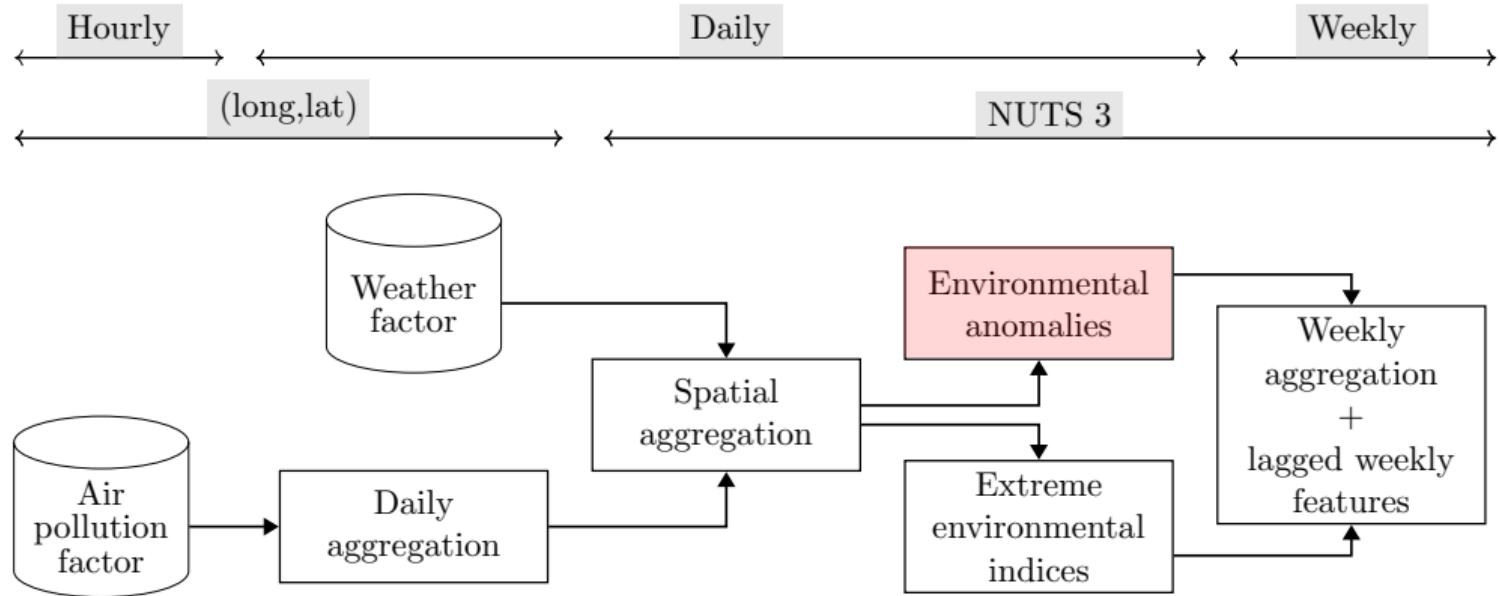
Define extreme high temperature index (hot-day index):

$$T.ind_{t,w,d}^{(r,95\%)} = \mathbb{1} \left\{ T_{\max,t,w,d}^{(r)} \geq q_{T_{\max}}^{(r,95\%)} \right\} + \mathbb{1} \left\{ T_{\text{avg},t,w,d}^{(r)} \geq q_{T_{\text{avg}}}^{(r,95\%)} \right\} + \mathbb{1} \left\{ T_{\min,t,w,d}^{(r)} \geq q_{T_{\min}}^{(r,95\%)} \right\}.$$

Index values: 0-3, indicating the severity of hot days.

Similar extreme indices are created for the other daily weather and air pollution factors.

Flow chart



Environmental anomalies

Create features that quantify deviations from typical, baseline conditions for each day throughout the year.

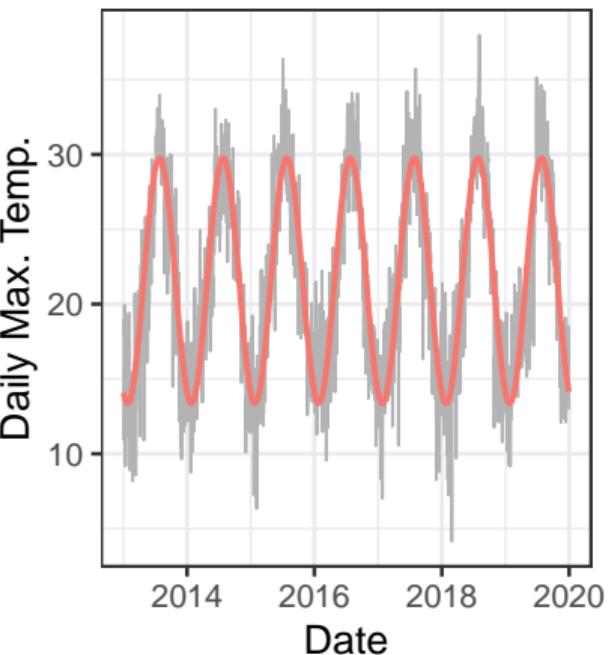
Robust linear regression to capture baseline:

$$\tilde{x}_{t,w,d}^{(r)} = \alpha_0^{(r)} + \alpha_1^{(r)} \sin\left(\frac{2\pi d}{365.25}\right) + \alpha_2^{(r)} \cos\left(\frac{2\pi d}{365.25}\right) + \epsilon_{t,w,d}^{(r)},$$

In the paper, we work with excesses or deviations from the baseline (anomalies):

$$\tilde{x}_{t,w,d}^{(r)} - \hat{x}_{t,w,d}^{(r)}$$

ES511: Barcelona



Environmental anomalies

Create features that [quantify deviations from typical, baseline conditions](#) for each day throughout the year.

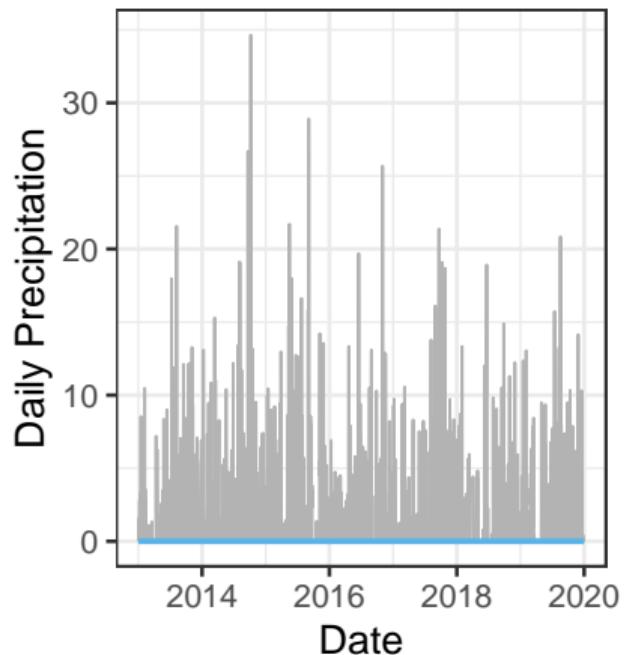
[Robust linear regression](#) to capture baseline:

$$\tilde{x}_{t,w,d}^{(r)} = \alpha_0^{(r)} + \alpha_1^{(r)} \sin\left(\frac{2\pi w}{365.25}\right) + \alpha_2^{(r)} \cos\left(\frac{2\pi w}{365.25}\right) + \epsilon_{t,w,d}^{(r)},$$

In the paper we work with excesses or deviations from the baseline ([anomalies](#)):

$$\tilde{x}_{t,w,d}^{(r)} - \hat{x}_{t,w,d}^{(r)}$$

SE110: Stockholms län



Accumulated local effects - motivation

How does each feature impact the predictions produced by the machine learning model?

Partial dependence plots interpret the marginal effect of a feature X_l on the model predictions as:

$$f_{l,\text{PDP}}(x) = \frac{1}{n} \sum_{j=1}^n \hat{f}(x, \mathbf{x}_j^{(-l)}),$$

where $\mathbf{x}_j^{(-l)}$ is the j -th training observation where the l -th feature is omitted.

Problem: correlated features may lead to unrealistic data points.

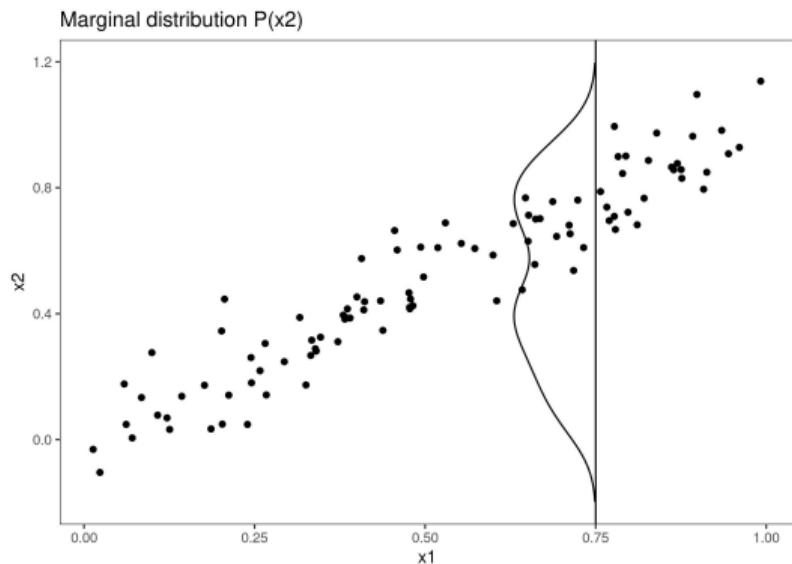


Figure: Visualisation taken from Molnar (2019).

Accumulated local effects - motivation

How does each feature impact the predictions generated by a machine learning model?

Partial dependence plots interpret the marginal effect of a feature X_l on the model predictions as:

$$f_{l,\text{PDP}}(x) = \frac{1}{n} \sum_{j=1}^n \hat{f}(x, \mathbf{x}_j^{(-l)}),$$

where $\mathbf{x}_j^{(-l)}$ is the j -th training observation where the l -th feature is omitted.

Problem: correlated features may lead to unrealistic data points.

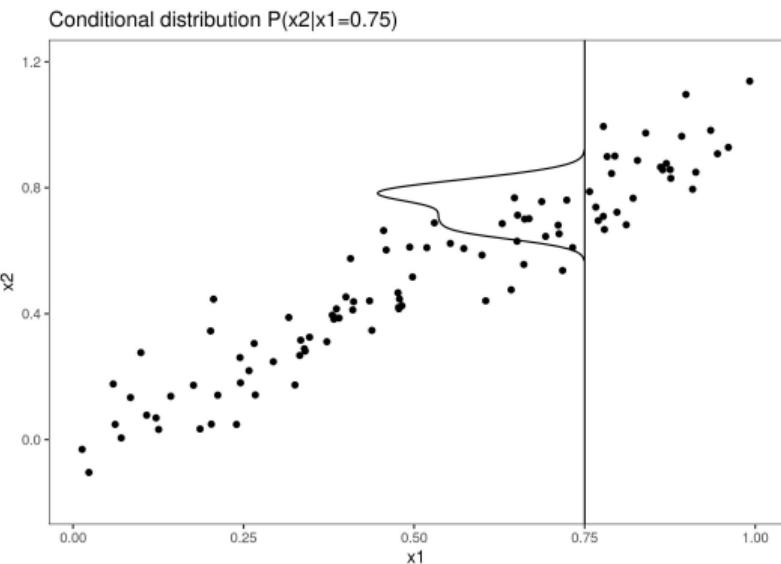


Figure: Visualisation taken from Molnar (2019).

Accumulated local effects - motivation

How does each feature impact the predictions generated by a machine learning model?

Partial dependence plots interpret the marginal effect of a feature X_l on the model predictions as:

$$f_{l,\text{PDP}}(x) = \frac{1}{n} \sum_{j=1}^n \hat{f}(x, \mathbf{x}_j^{(-l)}),$$

where $\mathbf{x}_j^{(-l)}$ is the j -th training observation where the l -th feature is omitted.

Problem: correlated features may lead to unrealistic data points.

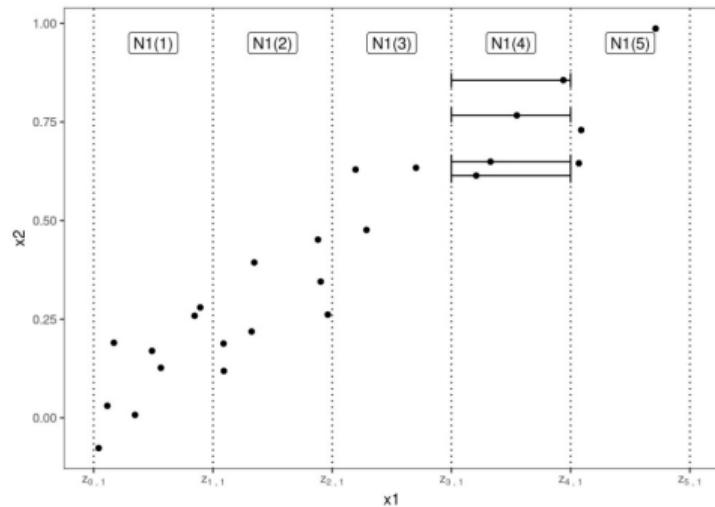
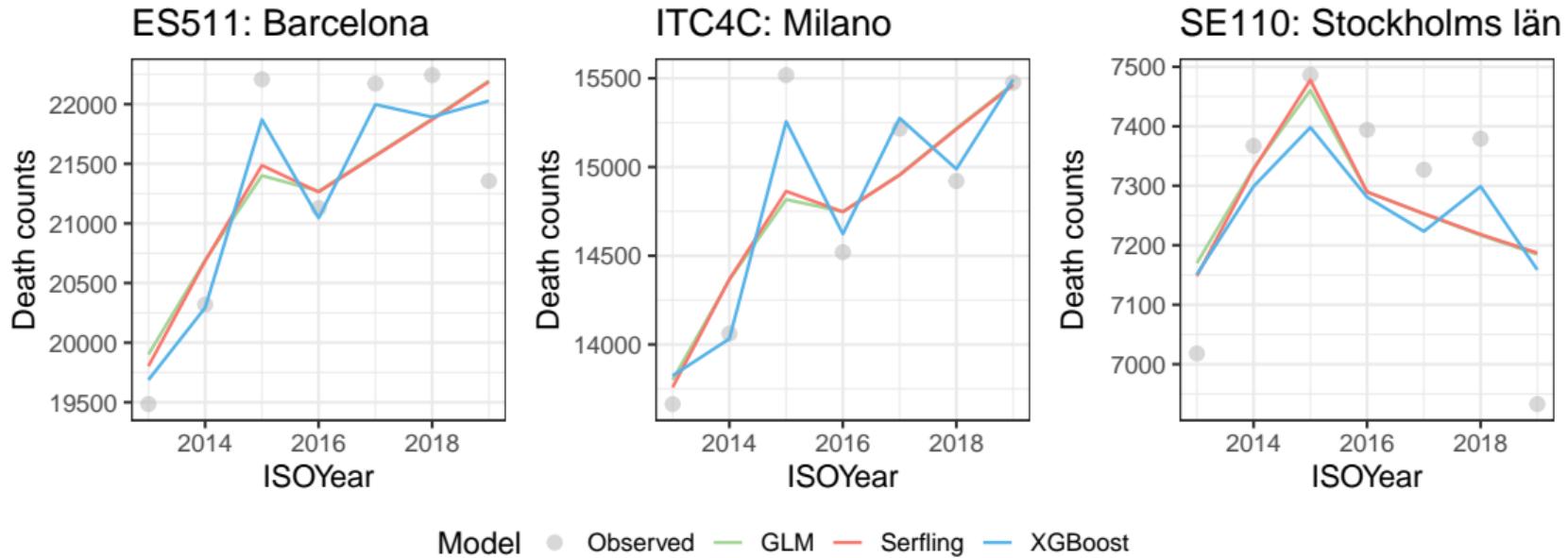


Figure: Visualisation taken from Molnar (2019).

Accumulated Local Effects avoid the use of unrealistic data instances, average changes in predictions and accumulate them. For a feature X_I , the ALE effect equals:

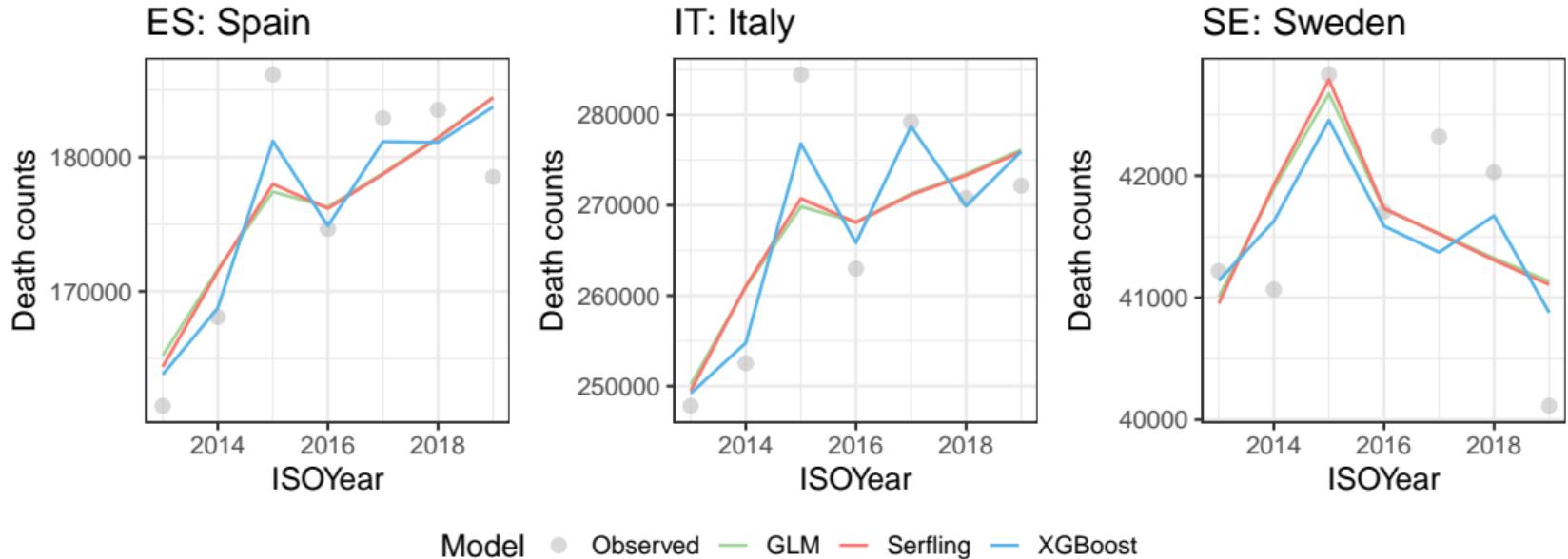
$$f_{I,\text{ALE}}(x) = \int_{z_{0,I}}^x \mathbb{E} \left[\frac{\partial f(X_1, X_2, \dots, X_p)}{\partial X_I} \middle| X_I = z_I \right] dz_I - c_I,$$

Instead of averaging predictions, ALE average changes of predictions (via the partial derivatives). The integral over z_I accumulates the differences in the predictions when moving over the range of X_I . A constant c_I is subtracted to center the ALE plot and make sure the mean effect is zero.



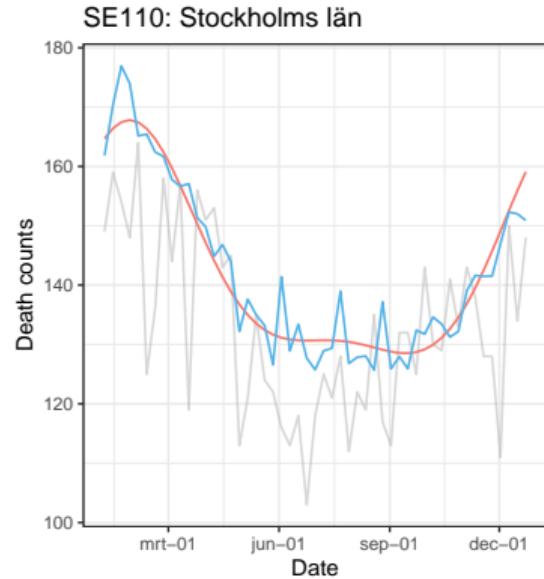
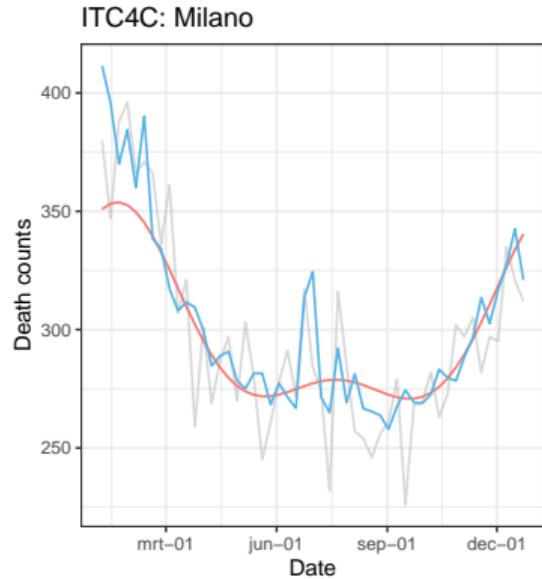
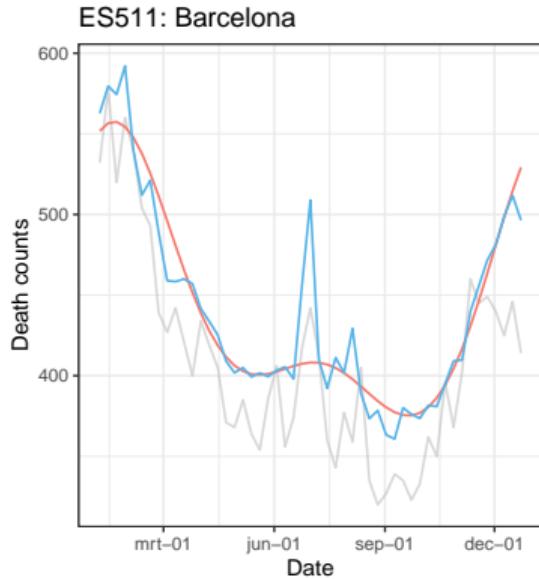
Spatial aggregation

40



Back-testing

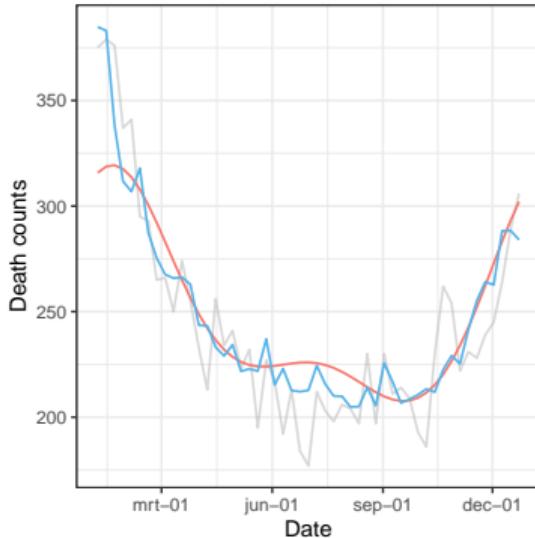
41



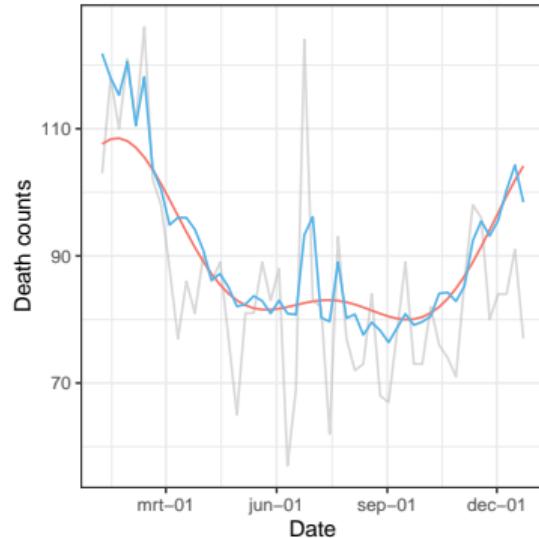
Back-testing

41

PT170: Área Metropolitana de Lisboa



ITC41: Varese



FRK26: Rhône

