EuroHPC Joint Undertaking

Rzeczpospolita Polska

The National Centre for Research and Development

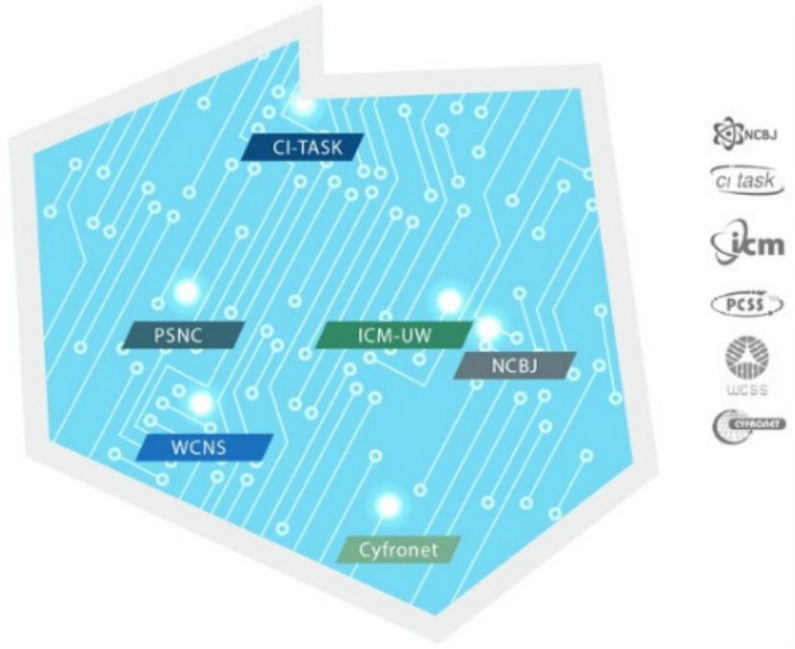
NATIONAL CENTRE FOR NUCLEAR RESEARCH ŚWIERK

EURO

„Machine learning for modeling mortality with respect to smog and ambient air temperature”

Polish National HPC Competence Centre

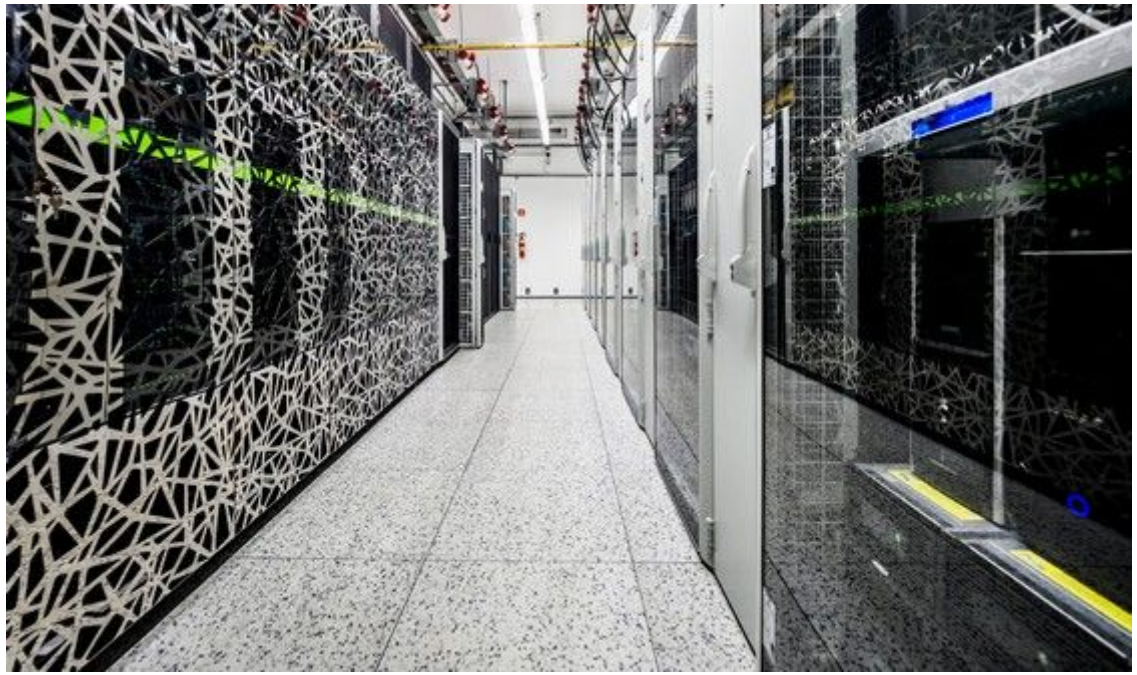
Polish NCC - collaboration of 6 biggest polish academic computing centers



Similar centers in all EU countries

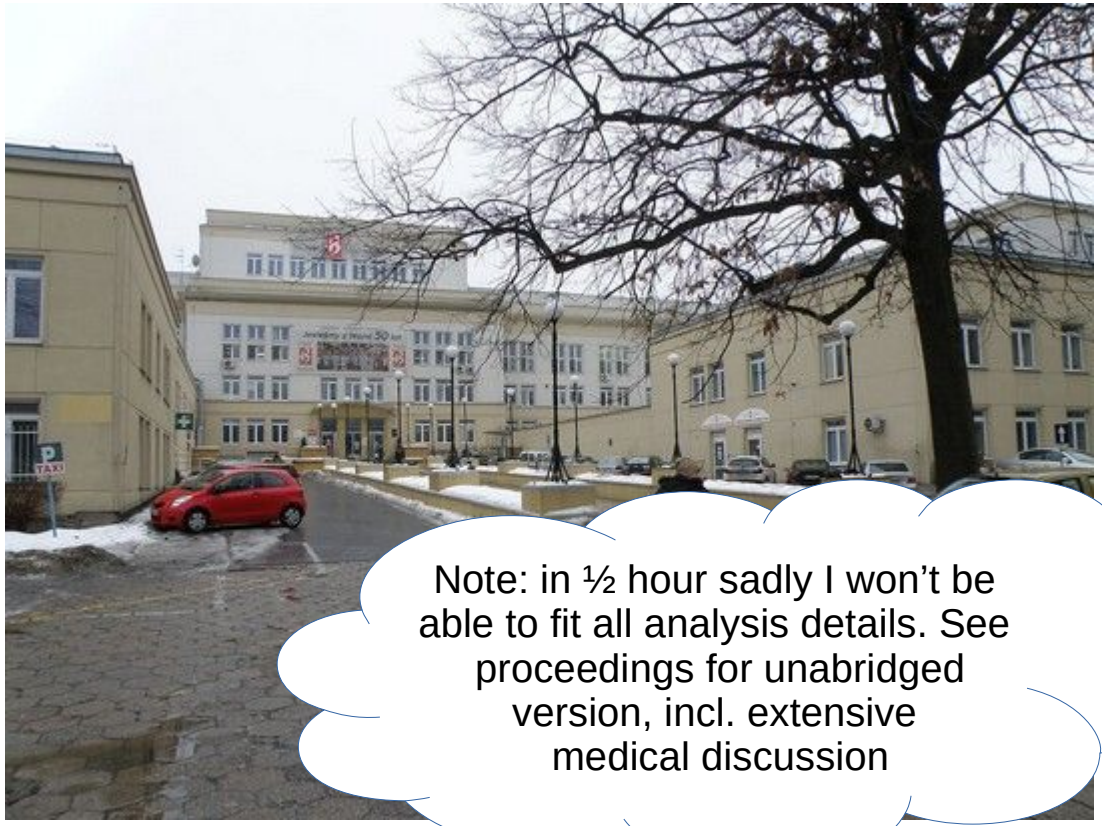
EuroCC project - goals

- To promote and provide support in:
  - High Performance Computing - **HPC**
  - Machine Learning - **ML**
  - High Performance Data Analitycs – **HPDA** (a.k.a BigData)
- Our mission is to help you (**free of charge**) in the above areas
  - support is targeted towards public institutions, academia and industry/SMEs
  - among others, it includes consulting or preparation of Proof-of-Concepts. **Ask me for more details during coffee break.**
- For additional info see also the **“EuroCC – National Competence Center for HPC”** talk tomorrow at 12:20



EuroCC project and this study

- Following slides present collaboration between experts of polish NCC and employees of Bielański Hospital in Warsaw
- We helped to analyze mortality data in order to determine if model linking levels of air pollution to number of deaths can be built
- Work of:
  - Krzysztof Nawrocki (NCBJ; ML expert)
  - Wojciech Moraczewski (anesthesiologist)
  - Prof. Marek Dąbrowski (cardiologist)
  - Dorota Gałczyńska-Zych (head of Bielański Hospital)
  - Tomasz Fruboes (NCBJ)



Note: in ½ hour sadly I won't be able to fit all analysis details. See proceedings for unabridged version, incl. extensive medical discussion

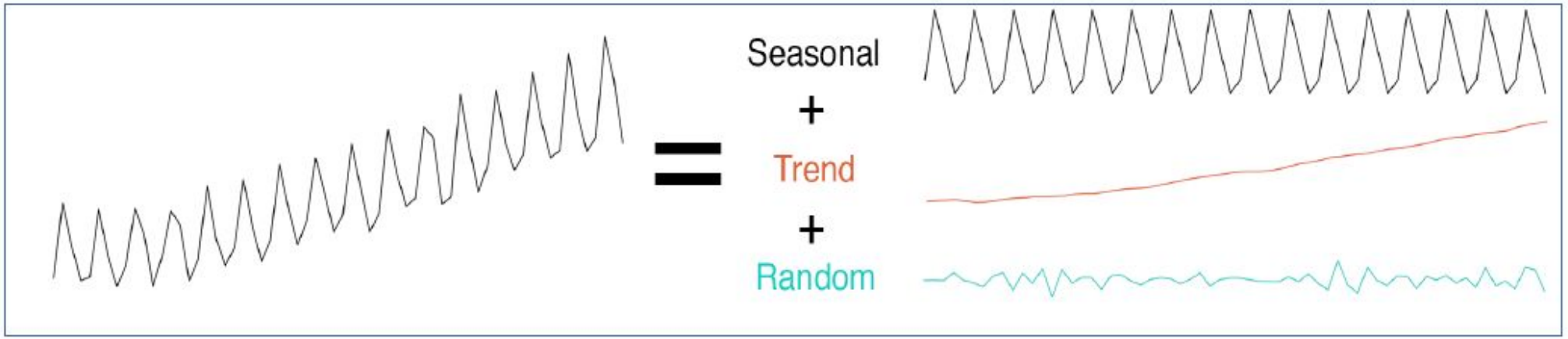
How it all started

- Polish National Health Fund (NFZ; polish agency financing public healthcare) noted, that in Jan 2017 number of deaths in Poland increased 23.5% wrt Jan 2016
- Analysis done by NFZ gave four possible reasons:
  - Low ambient air temperature
  - Record high air pollution
  - Flu epidemic
  - Changes in the rules for reimbursement of treatments of acute coronary syndrome
- We decided to check if this effect would be visible on a single-hospital level and what we can learn from data present in Hospital Information System
  - data anonymized prior to the analysis by hospital staff

Jacek Halicki, CC SA 4.0

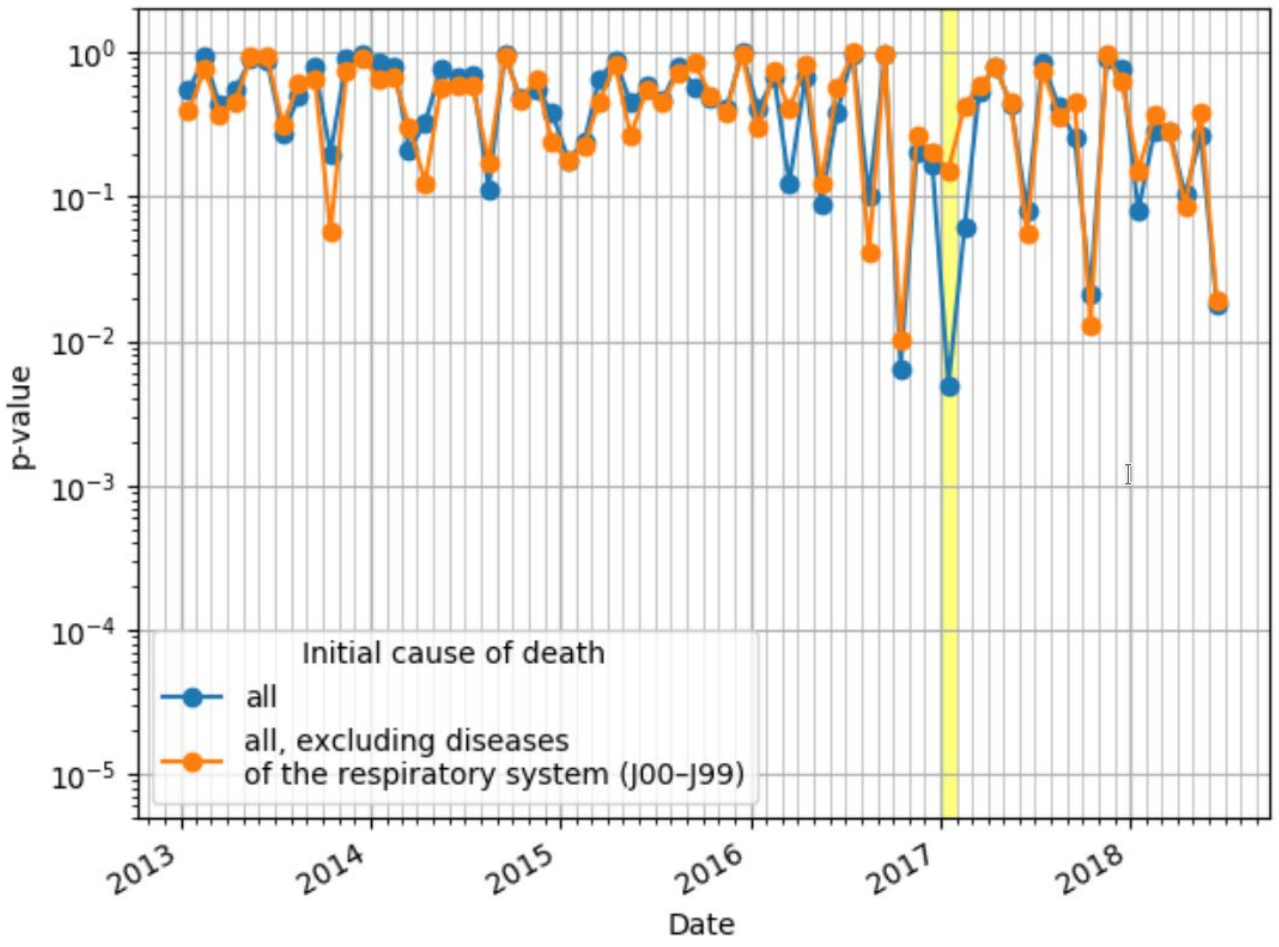
Do we observe mortality excess for hospital data?

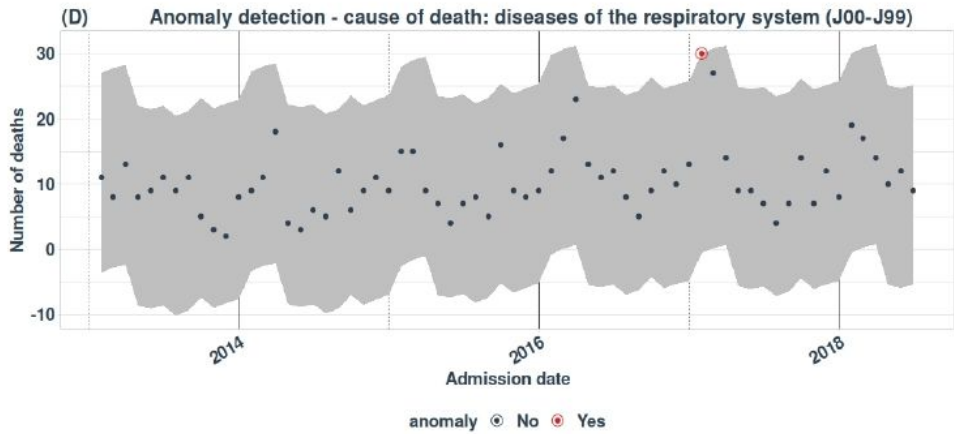
- Analyze 2014-2018 data. Check number of deaths in every month (time series analysis)
- Two approaches followed:
  - With R Anomalize library - decompose time series into seasonal changes/trend/random fluctuations. Check if given point did not fluctuate to far wrt seasonal+trend baseline (use Generalized ESD Test for Outliers)
  - With permutation test (exact test; no assumptions on data distribution) – how different is given month (e.g. Jul'15) wrt to same period in other years (Jul'14, Jul'16,Jul'17, Jul'18)

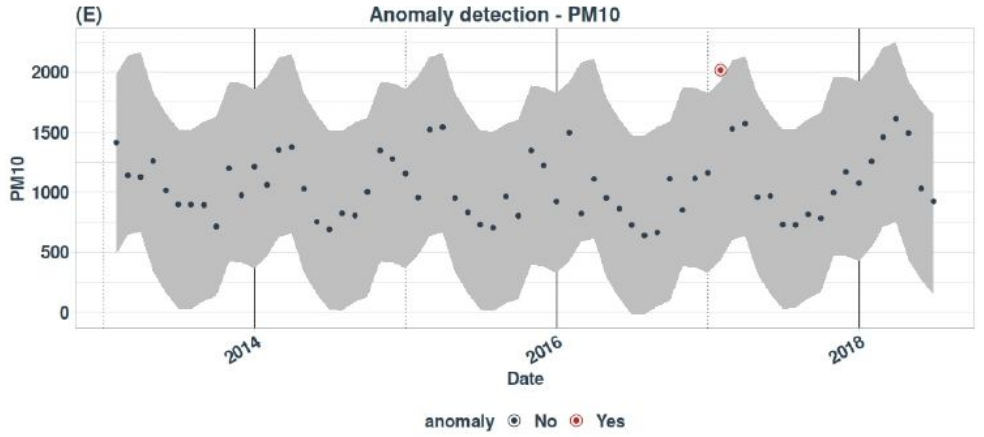


With both approaches we see anomaly for Jan 2017

Alpha = 0.0001
Time decomposition method = STL
Remainder anomalize method = GESD



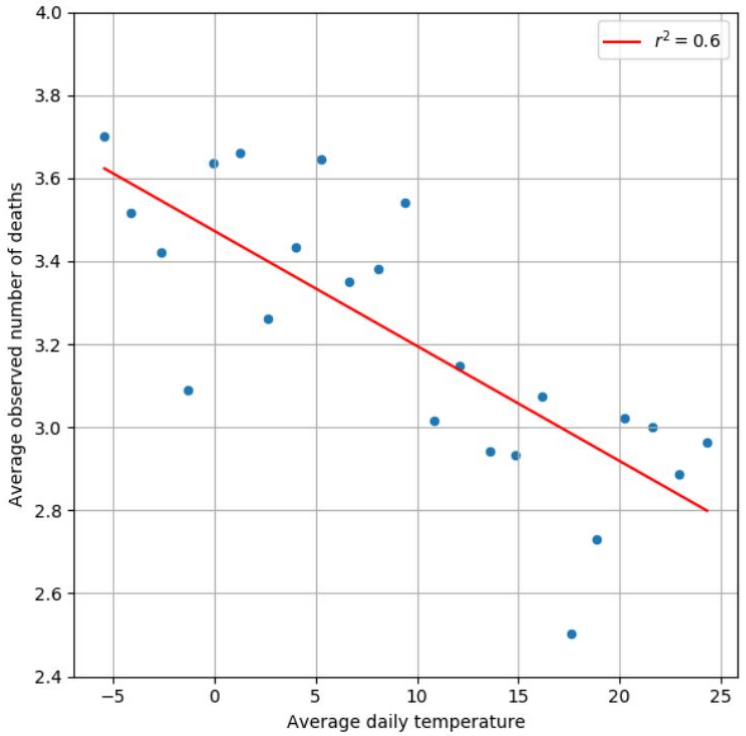




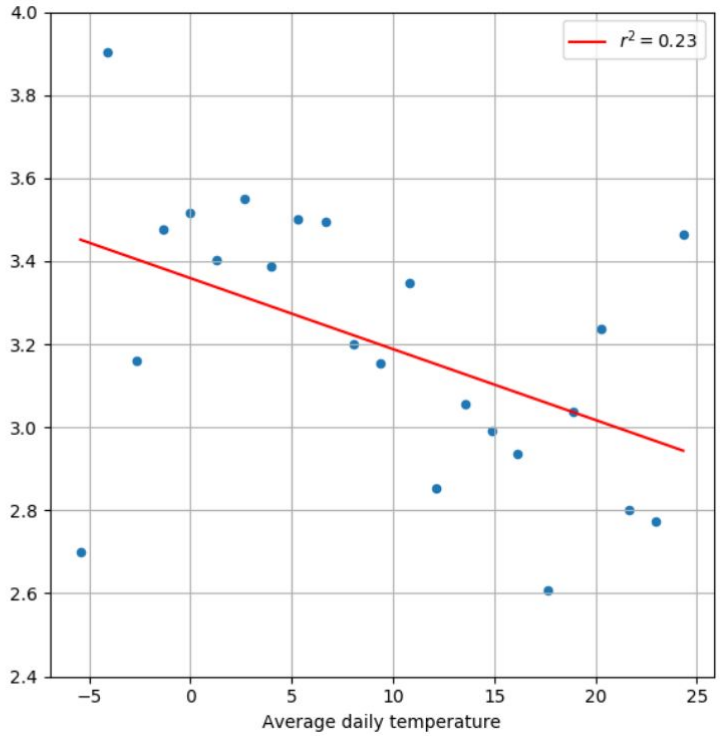
EuroHPC Joint Undertaking
Rzeczpospolita Polska
The National Centre for Research and Development

Building mortality model – can we explain the Jan 2017 excess with ML?

- Mortality depends on ambient air temperature prior to date of hospital admission
- Taking ambient air temperature prior to day of death makes no sense, as patient may stay in Intensive Care for a long period of time



Air temp. prior to admission

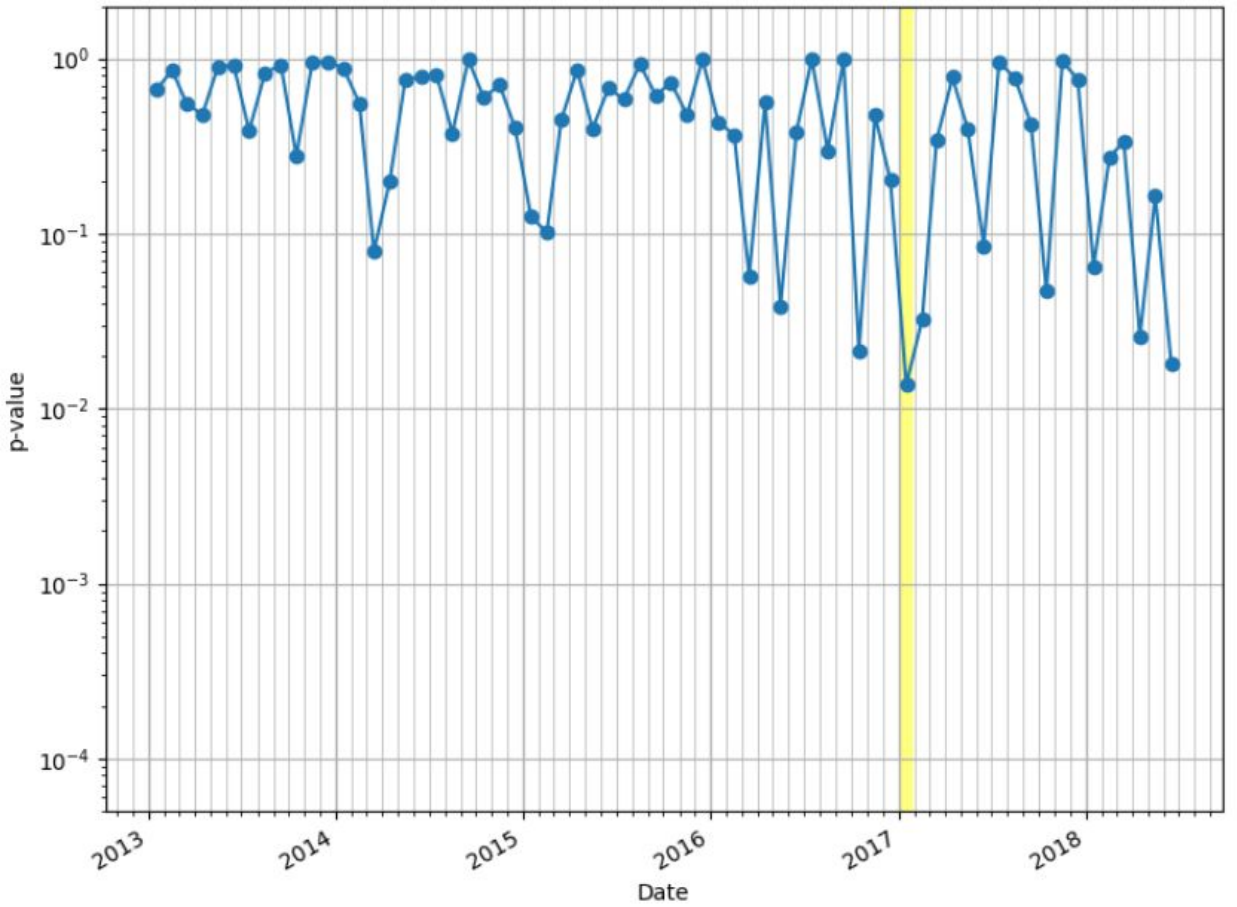


Air temp. prior to death

EuroHPC Joint Undertaking
Rzeczpospolita Polska
The National Centre for Research and Development

Mortality model

- We modell daily expected number of deaths as a function of average ambient air temperature
- Observed number of deaths follows Poisson distribution with the above expected value
- With such approach we are able to obtain p-value for every month – probability of obtaining observed or larger number of deaths wrt model prediction
  - The lower p-value - the more anomalous is given month
- **Obtained model (based only on ambient air temperature) is not able to explain the Jan 2017 excess**



EuroHPC Joint Undertaking
Rzeczpospolita Polska
The National Centre for Research and Development

New input variables

- For next iteration we added variable corresponding to air pollution levels (PM10 – amount of inhalable particles, with diameters of 10 micrometers and smaller; measured in $\mu g/m^3$)
- We also tested multiple ways of input variables averaging:
  - Different time offset w.r.t. admission date
  - Different window size for averaging
  - For air pollution input variable we also tested if discriminating input data against a threshold of 50 $\mu g/m^3$ (i.e. current `safe` level, specified by polish environment agency) leads to a better model

Above leads to large number of models. How to pick best one?

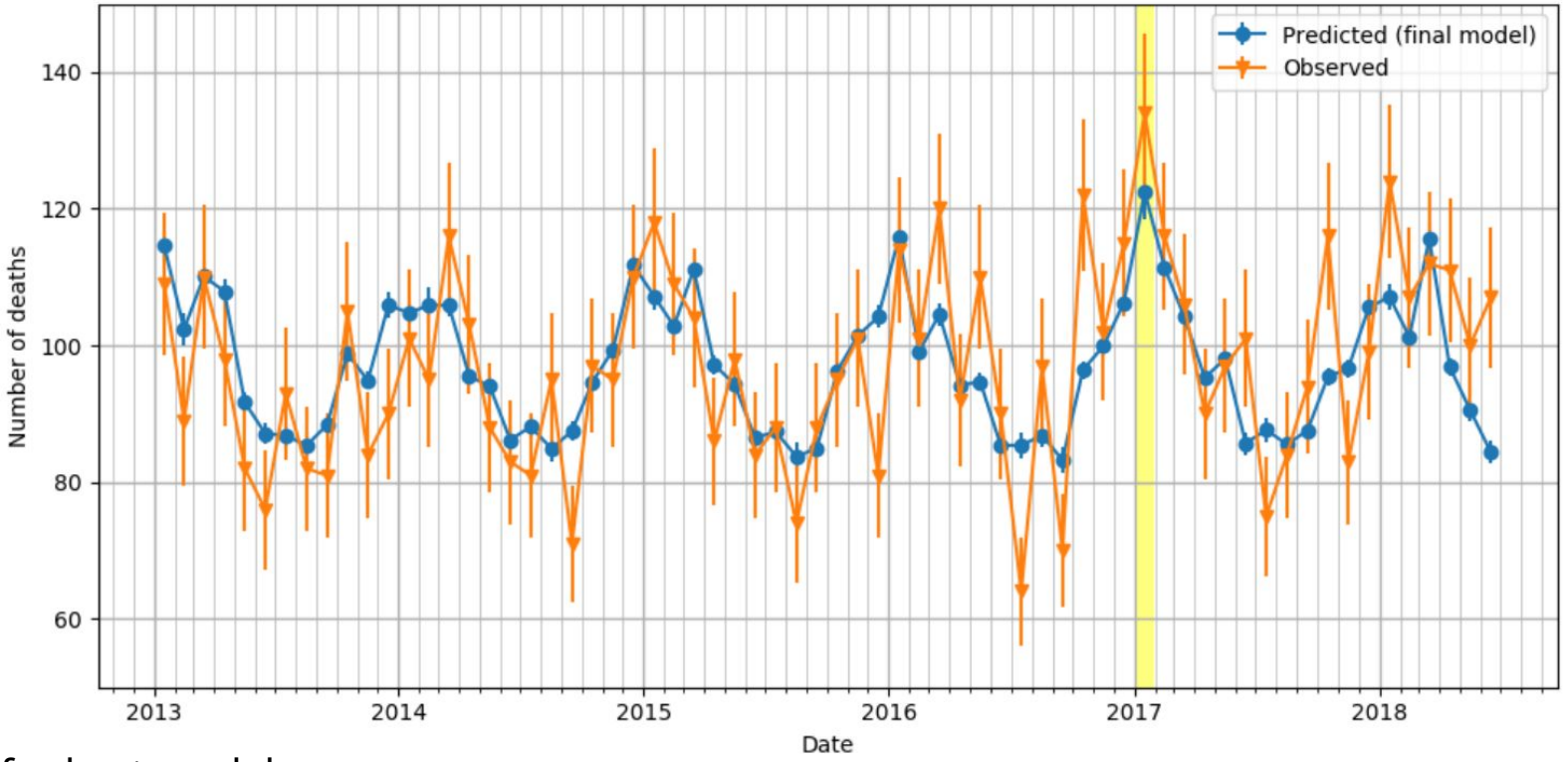
- Typically, one will use k-fold cross validation
  - Penalty – longer computation time (equal to number of folds)
- For linear models, alternative exist!

Variable selection procedure

Best input variables selected by comparing models using Akaike information criterion (AIC):

- Method (asymptotically) equivalent to k-fold cross validation (for linear models)
- **Penalty for each additional input variable in model**
- **Single fit (whole dataset) per tested set of variables**
- Allows for quantitative models comparison. Procedure:
  - Compare some model (“model B”) to the best performing model found (“model A”)
  - Result - P^{AIC} := probability, that model B in fact describes data better than model A

Mortality model



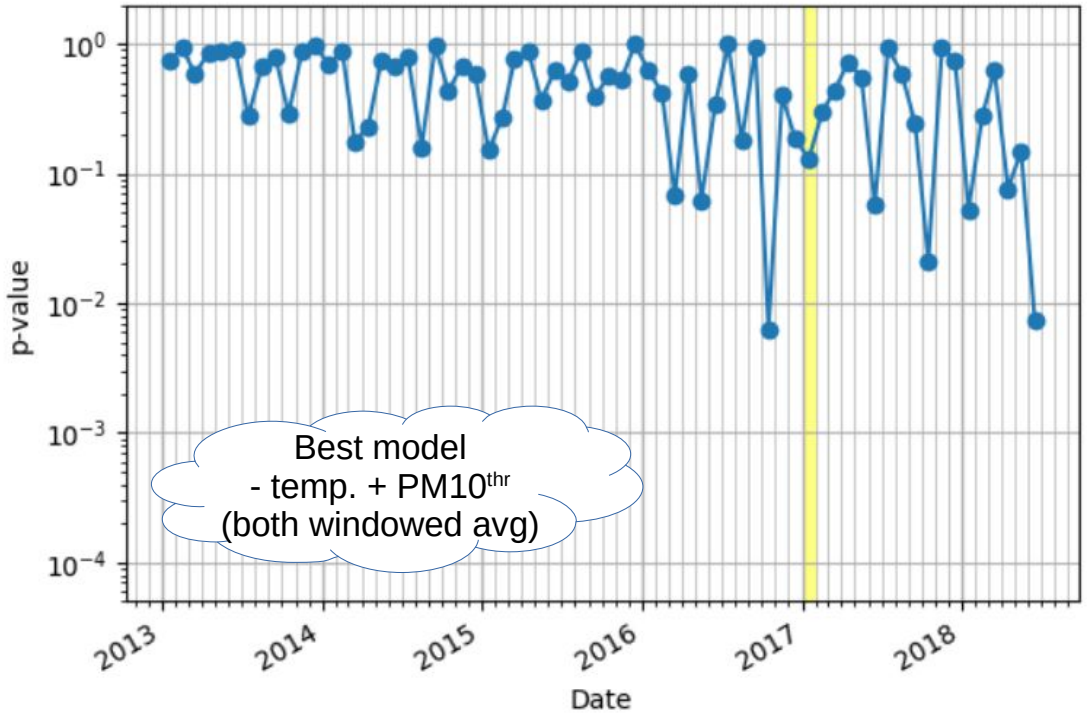
Input variables selected for best model:

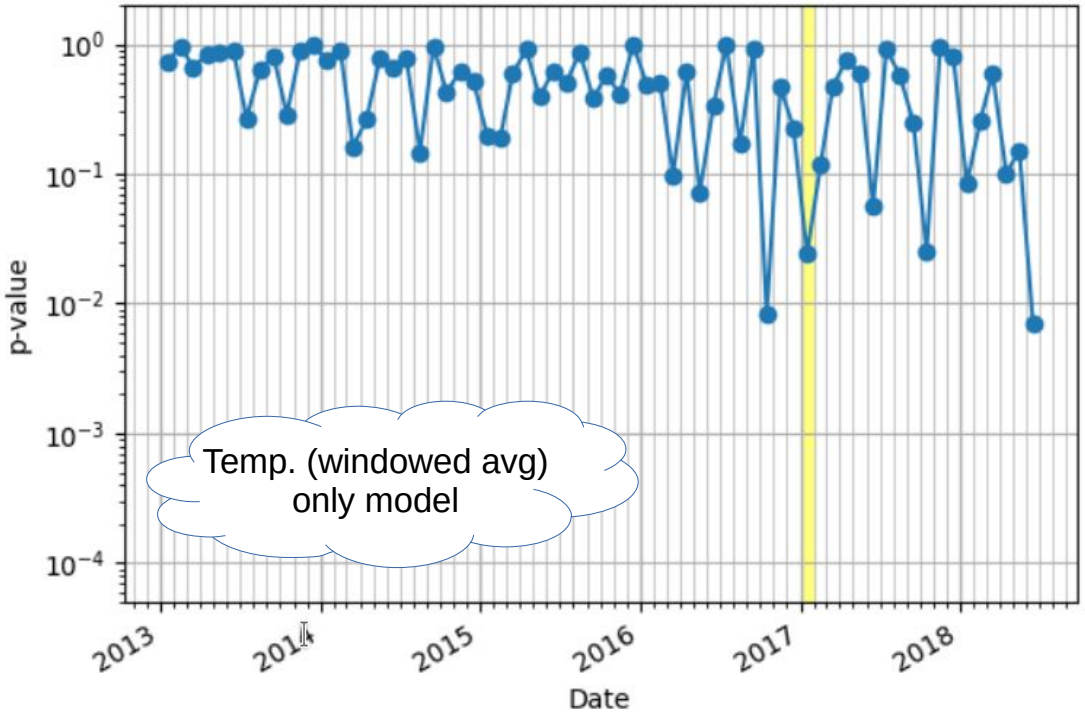
- windowed average ambient air temperature – days 17...13 prior to admission
- windowed average of PM10 level – days 8...1 prior to admission, with 50 $\mu g/m^3$ threshold

Second best model was temperature only model, with $P^{AIC} = 0.06$ - discarded

EuroHPC Joint Undertaking
Rzeczpospolita Polska
The National Centre for Research and Development

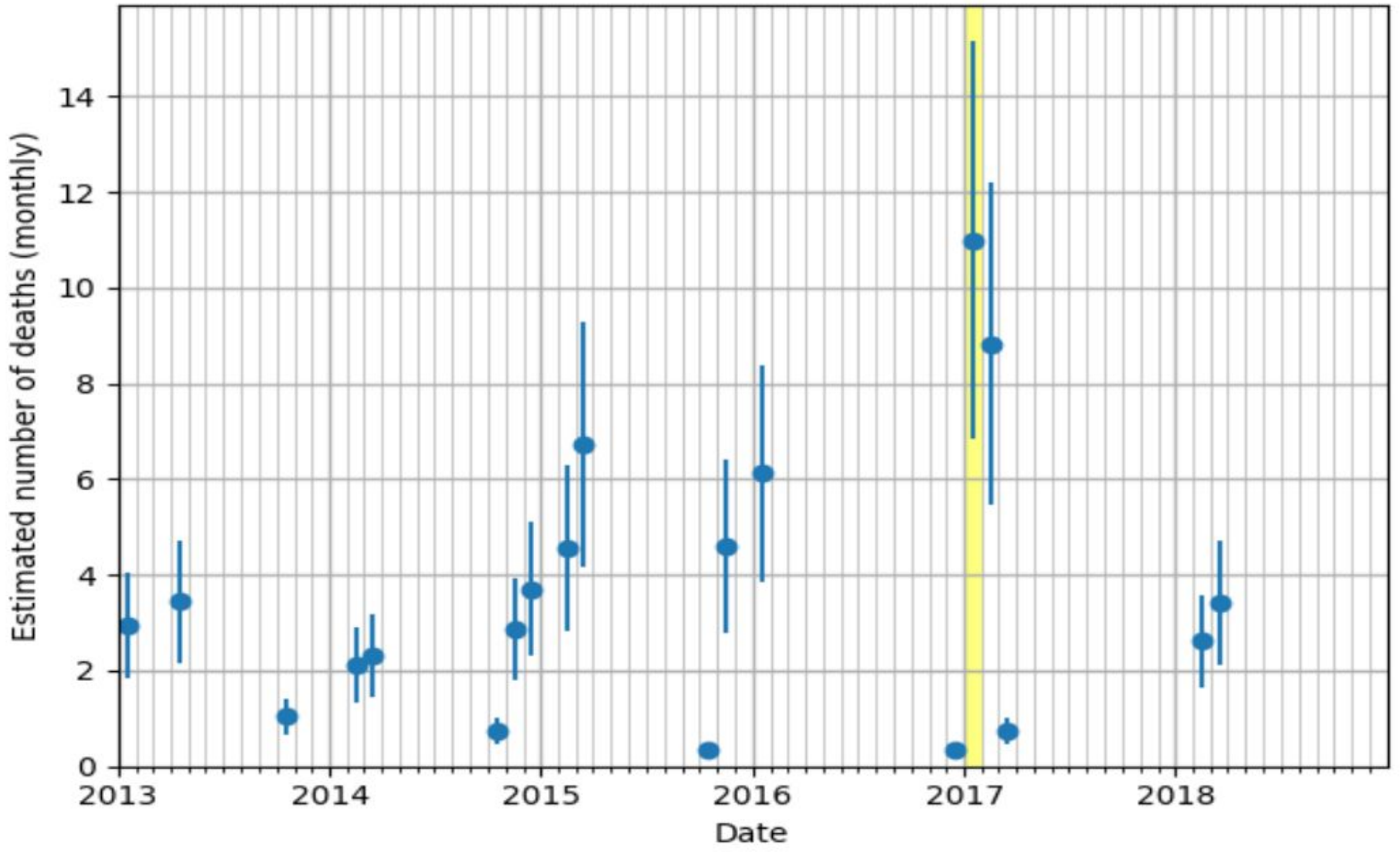
P-values comparison – temp+PM10 vs temp-only model





EuroHPC Joint Undertaking
Rzeczpospolita Polska
The National Centre for Research and Development

Number of deaths attributed to air pollution



EuroHPC Joint Undertaking
Rzeczpospolita Polska
The National Centre for Research and Development

Summary

- Number of deaths observed in Jan 2017 in Bielański Hospital is anomalous (consistent with country-wide observation)
  - Excess due to respiratory-related deaths
- Mortality model utilizing ambient air temperature and air pollution as input variables is able to explain the excess
- **Model attributes 8% of deaths observed in Jan 2017 to air pollution**

See proceedings for a more detailed version, incl. extensive medical discussion

EuroHPC Joint Undertaking
Rzeczpospolita Polska
The National Centre for Research and Development

Thank you!

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 951732. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, United Kingdom, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Switzerland, Turkey, Republic of North Macedonia, Iceland, Montenegro