Smog prediction

Data Analytics

Composition of the team

- Joanna Nużka, 400561
- Katarzyna Słomińska, 400563

Introduction

The goal of our project is to predict amount of smog based on weather and time related data like day of week or heating season. We are analyzing one of smog indicators – PM10. It is a composition of molecules with maximal diameter 10µm and it is one of the major indicators used in Poland.

Dataset

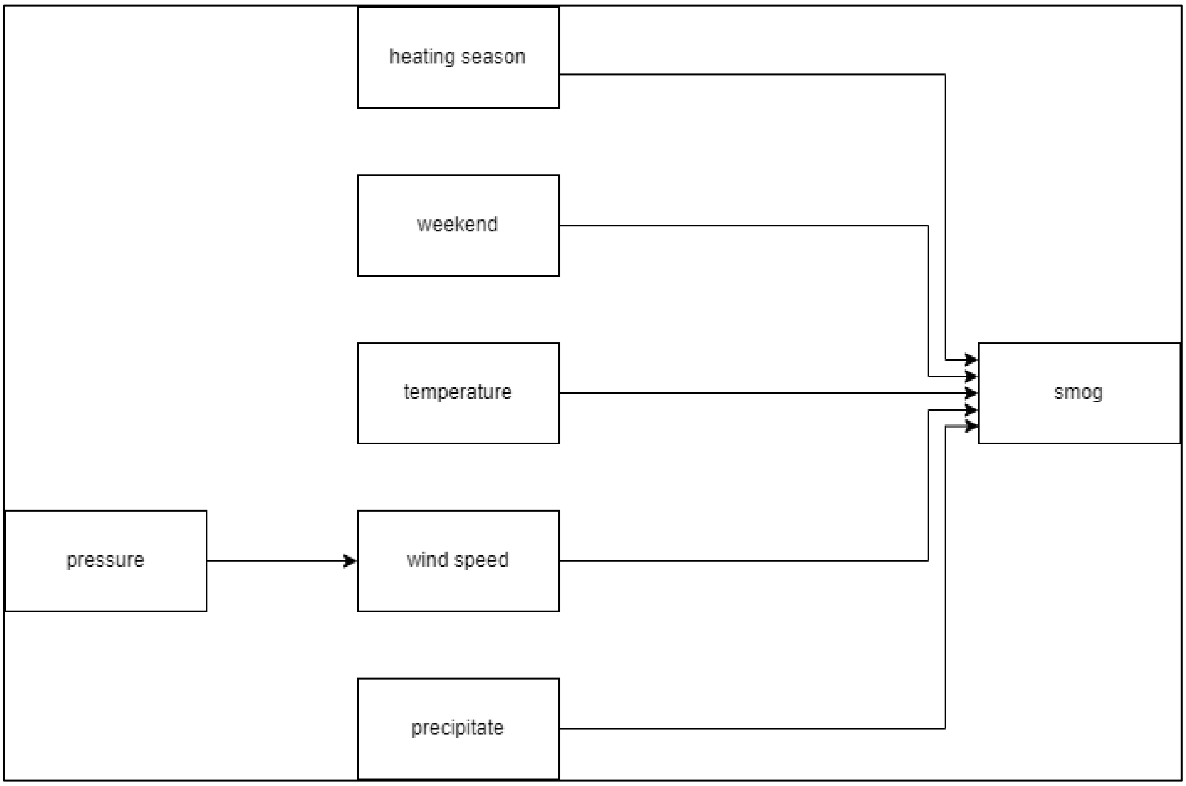
We use data from few sources: weather data from Visual crossing, smog data from powietrze.gios.gov.pl and time related data from calendar. We chose some of available data and made dataset that contains:

- date,
- temperature,
- wind speed,
- appearance of precipitate,
- pressure,
- if there was a heating season,
- if there was a weekend,
- smog measured on Złoty Róg station,
- smog measured on Bulwarowa station,
- smog measured on Swoszowice station.

We chose station which are characterized by different locations and environment. Swoszowice is background station located in less urbanized environment, near the green areas. Złoty Róg Street is background station too but it lies closer to the city center in area with bigger traffic movement. Bulwarowa Street is industrial station so it can measure the biggest amount of smog. The location of these stations is dictated by the choice of places with different traffic volumes

DAG

The DAG diagram is shown below. This is a simple diagram because we are analyzing the occurrence of smog at three different stations. Therefore, make separate models for each station. All the necessary weather data is in the datasets we have, which reduced the need to model the paramters.



Confoundings

Forks

We don't have forks in our model.

Colliders

- Wind speed depends on pressure.
- Smog depends on heating season, weekend, temperature, wind speed and precipitate.

Pipes

- Pressure is transmitted through wind speed to smog.

Data preprocessing

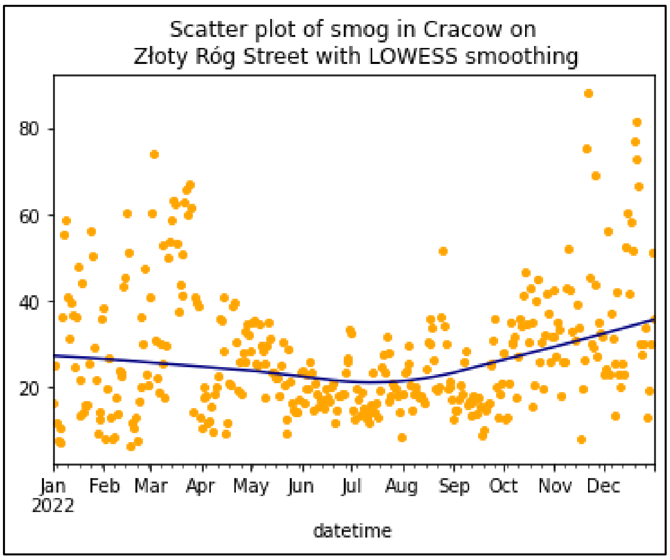
To prepare the data for modeling and make it more manageable, we took the following steps:

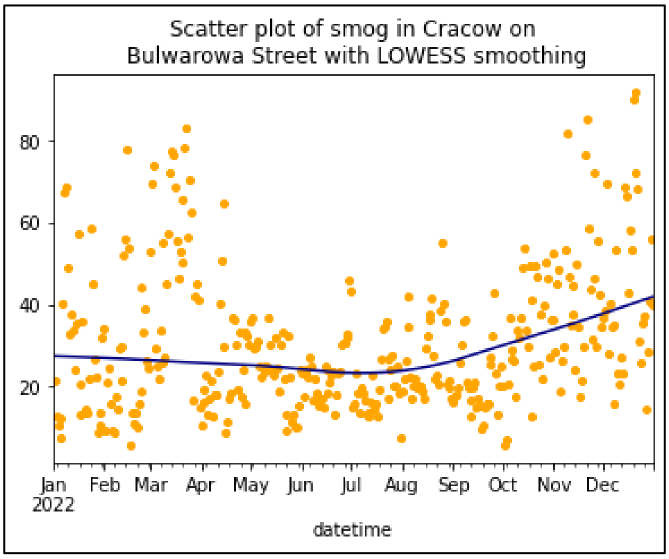
1) **Downloading data** – we downloaded smog intensity and weather data
2) **Cleaning the data –** we removed data that was not needed in our project, which helped to organize it
3) **Sorting the data –** we sorted the two databases by date so that in later steps we could merge them together so that the date of the weather phenomenon coincided with the date of the smog occurrence
4) **Extracting the data set –** we then prepared a data set for the amount of smog recorded at three different research stations. We take into account three different research stations because of the recorded data gaps on some days. This will allow us to aggregate better results than if we used the average value from all stations
5) **Data removal** – we remove zero values from the smog data set

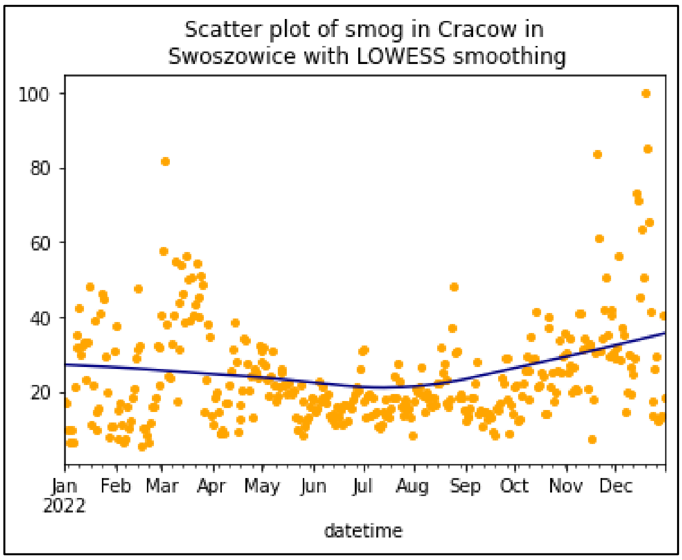
   Pre-processing of data – we perform pre-processing of the data before calling the state models: we calculate the difference between the pressure and its average value, in order to use it to predict the wind speed based on it

Data visualization

Scatter plots of smog data in time visualization:

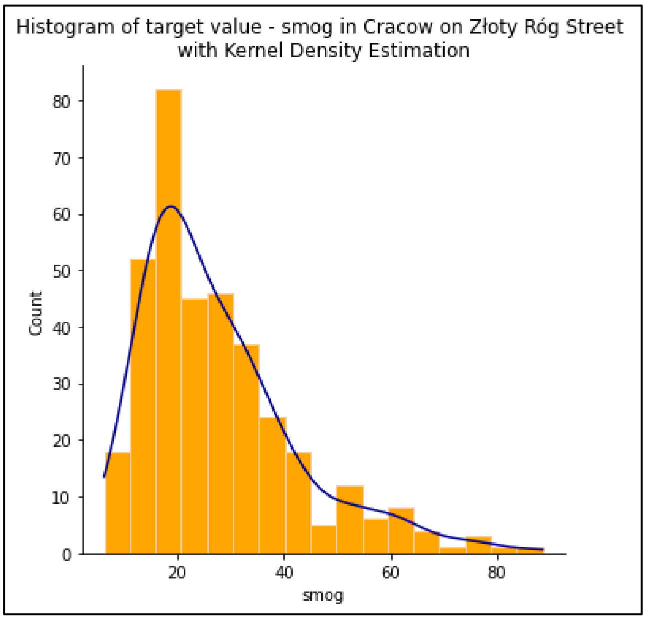


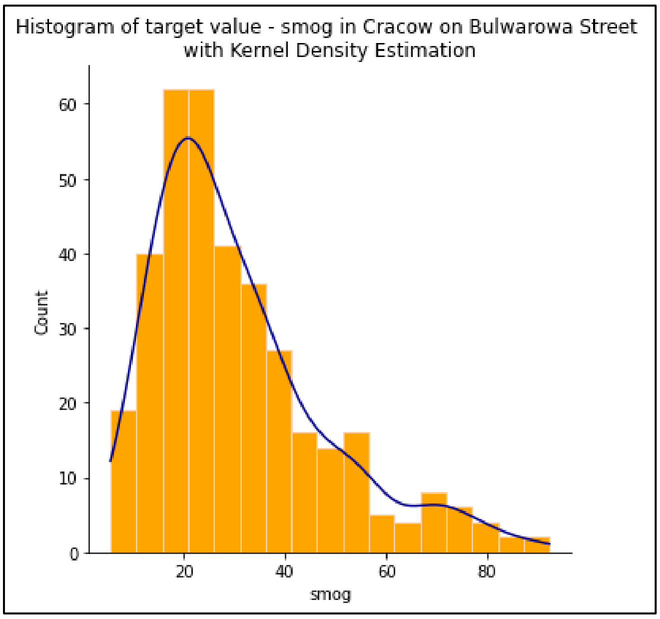


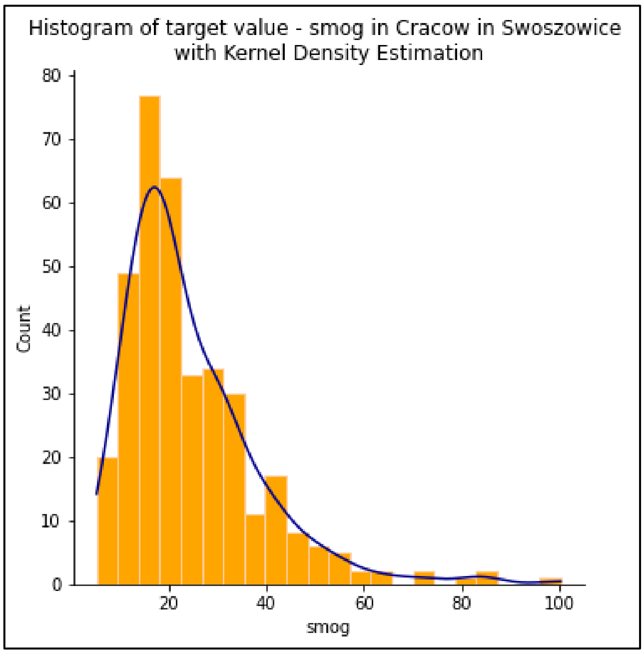


Histogram plots of smog data:

Smog data are characterized by big variety from LOWESS smoothing line.







In every station smog data have slightly different distribution so we have to create separate models for each of them.

Models

For every measurement station we use two linear regression models: one with normal distribution and second one with student's t-distribution. We want to compare their performance in our problem. One of the student's t-distribution parameters is number of degrees of freedom. If this number is bigger than 30 then the student's t-distribution is very similar to normal distribution. We chose 5 degrees of freedom so our student's t-distribution lower peak and higher tails than normal distribution. Smog data are characterized by big variety so student's t-distribution may perform better than normal.

Sampling was successful in both models

Inputs for every model

- **N** – number of samples,
- **precipprob** – vector with precipitate (values 1 if appeared or 0 if not),
- **wind_speed** – vector with wind speed values,
- **heating_sezon** – vector with information about heating season (1 if were or 0 if not),
- **weekend** – vector with information about weekend (1 if were or 0 if not),
- **temp** – vector with temperature values,
- **smog** – vector with smog values.

Parameters for every model

- **alpha, beta_temp, beta_ws, d_pp, d_hs, d_wn** – coefficients used to count mean of distributions,
- **sigma** – standard deviation of distributions,
- **mu** – mean of distributions.

Formulas

Złoty Róg station:

alpha ~ normal(30, 2);
beta_temp ~ normal(0, 1);
beta_ws ~ normal(0, 1);
d_pp ~ normal(-9, 1);
d_hs ~ normal(13, 1);
d_wn ~ normal(1, 2);
sigma ~ normal(15, 1);
*mu = alpha + beta_temp * temp + beta_ws * wind_speed + d_pp * precipprob + d_hs * heating_sezon + d_wn * weekend;*

Normal distribution:
smog ~ normal(mu, sigma);

Student's t-distribution:
smog ~ student_t(mu, sigma);

Bulwarowa station

alpha ~ normal(30, 2);
beta_temp ~ normal(0, 1);
beta_ws ~ normal(0, 1);
d_pp ~ normal(-12, 1);
d_hs ~ normal(13, 1);
d_wn ~ normal(-1, 2);
sigma ~ normal(15, 1);
*mu = alpha + beta_temp * temp + beta_ws * wind_speed + d_pp * precipprob + d_hs * heating_sezon + d_wn * weekend;*

Normal distribution:
smog ~ normal(mu, sigma);

Student's t-distribution:
smog ~ student_t(mu, sigma);

Swoszowice station

alpha ~ normal(30, 2);
beta_temp ~ normal(0, 1);
beta_ws ~ normal(0, 1);
d_pp ~ normal(-6, 1);
d_hs ~ normal(11, 1);
d_wn ~ normal(2, 2);
sigma ~ normal(15, 1);
*mu = alpha + beta_temp * temp + beta_ws * wind_speed + d_pp * precipprob + d_hs * heating_sezon + d_wn * weekend;*

Normal distribution:
smog ~ normal(mu, sigma);

Student's t-distribution:
smog ~ student_t(mu, sigma);

We also try to use different model where wind speed isn't an input but parameter given by formula:

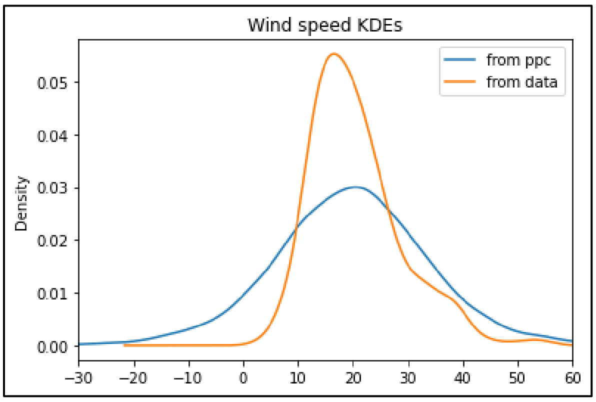
input: pressure

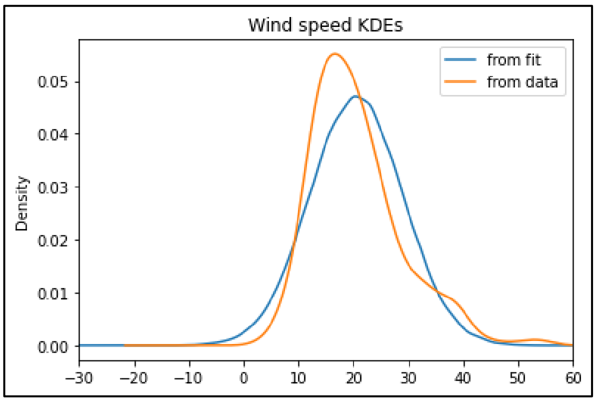
alpha_wind ~ normal(20, 1);
beta_wind ~ normal(0, 1);
sigma_wind ~ normal(10, 1);
*wind_speed ~ normal(alpha_wind + beta_wind * pressure, sigma_wind);*

We want to build model based on DAG diagram and compare performance of models

Model of wind speed from pressure:

In real life wind speed highly depends on pressure so we try to model it. We receive such results:





We received some negative results in KDE for wind speed. It is acceptable because in prior analysis we use a normal distribution to model the pressure so it can have values differents than are in nature. While fitting model to real pressure data we received similar density to this from data. We use this relationship to find formula for modeling wind speed based on pressure.

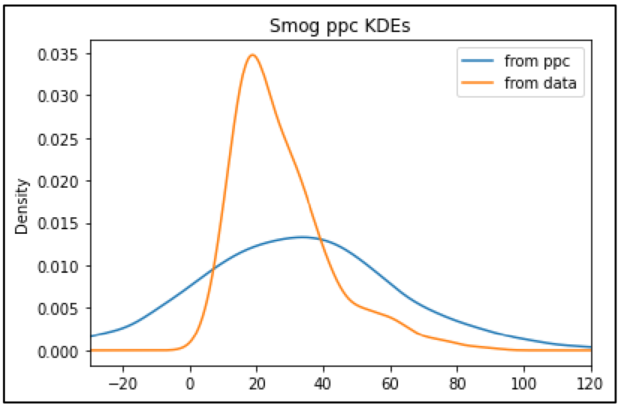
Priors

We use below formulas to model the data:

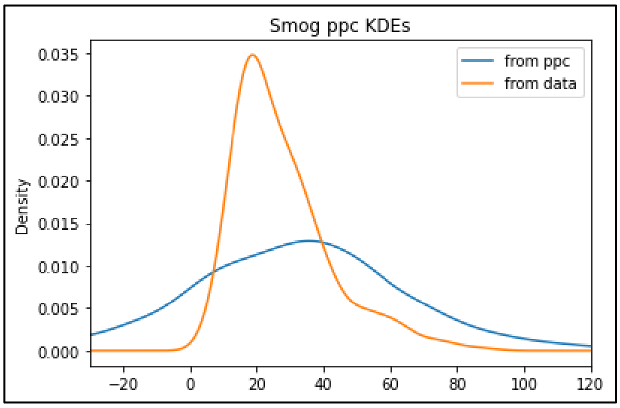
> *windspeed = normal_rng(20, 10);*
> *precipprob = bernoulli_rng(0.67);*
> *precipprob = bernoulli_rng(0.67);*
> *weekend = bernoulli_rng(0.14);*
> *temp = normal_rng(10, 8);*

Then we model smog based on the formulas described before.

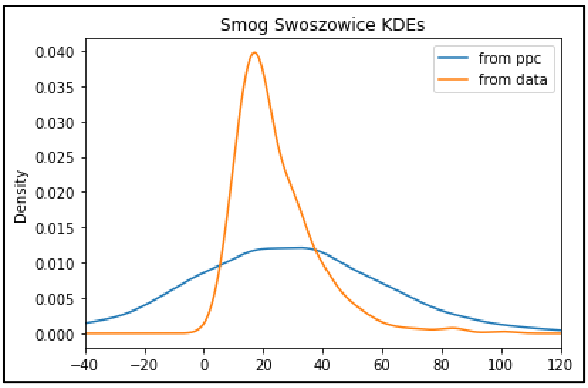
Złoty Róg station and normal distribution:



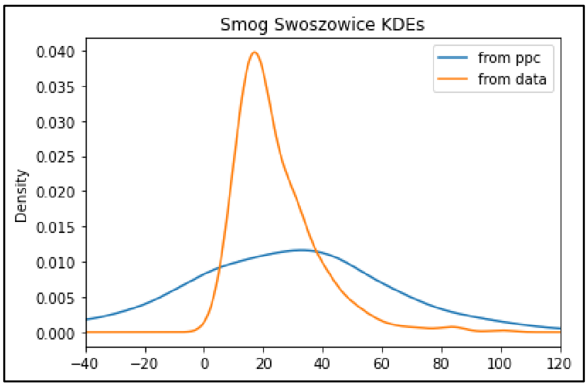
Złoty Róg station and t_student distribution:



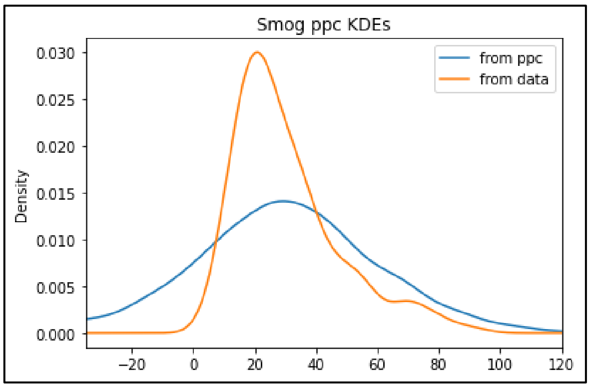
Bulwarowa station and normal distribution:



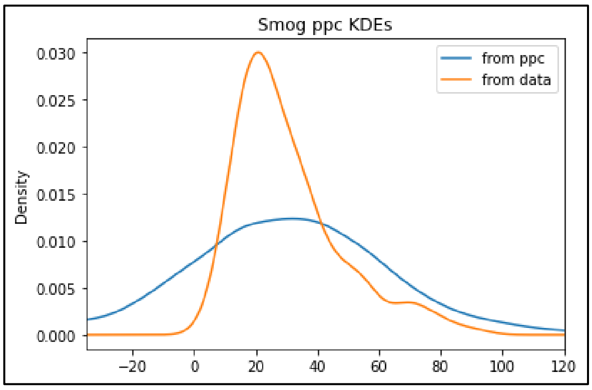
Bulwarowa station and t_student's distribution:



Swoszowice station and normal distribution:



Swoszowice station and t_student's distribution



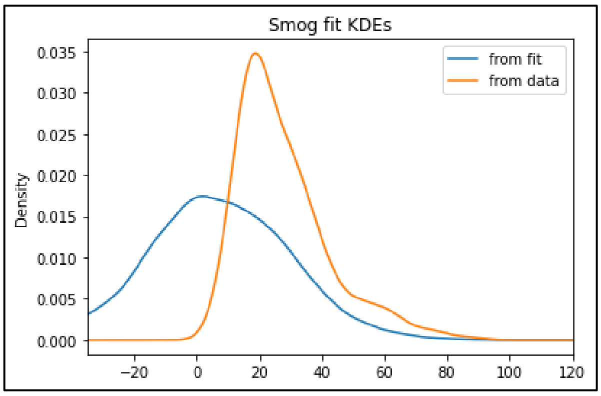
KDEs plots are different for smog received from model and from data. That's because in model we have distributions not real data so results also may be different. But we can see that in both cases densities have peaks in similar values.

Posteriors

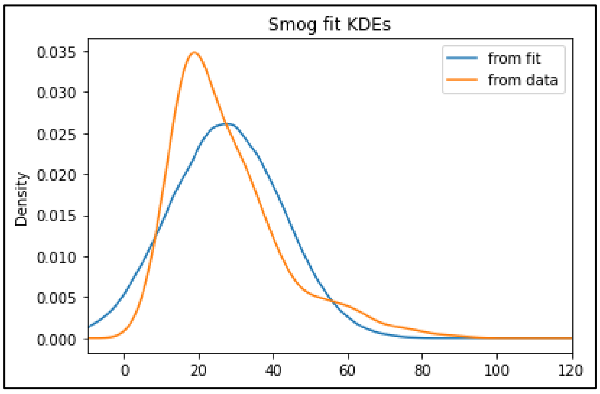
We create posteriors models based on the formulas described before.

Złoty Róg station:

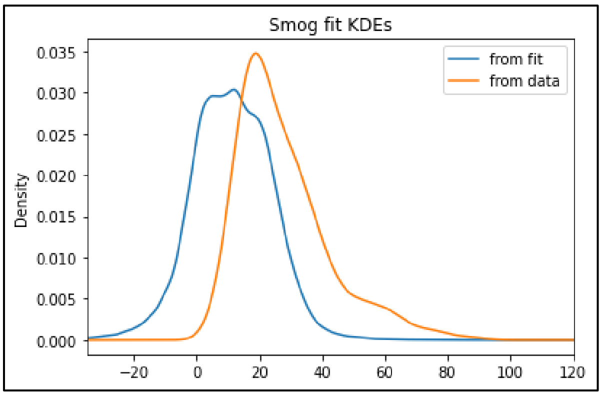
Normal distribution with wind speed modeled:



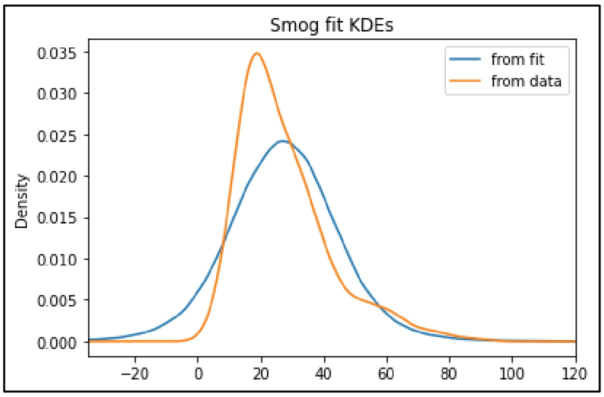
Normal distribution with wind speed input:



Student's t-distribution with wind-speed modeled:



Student's t-distribution with wind speed input:

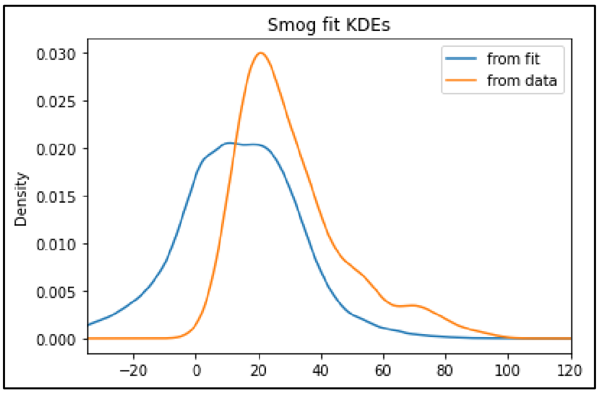


Smog data have values concentrated in one place but they aren't symmetric. Normal and student's t-distribution are symmetric and for models with wind speed input they have tails in the same places as data but their peaks are quite lower and shifted.

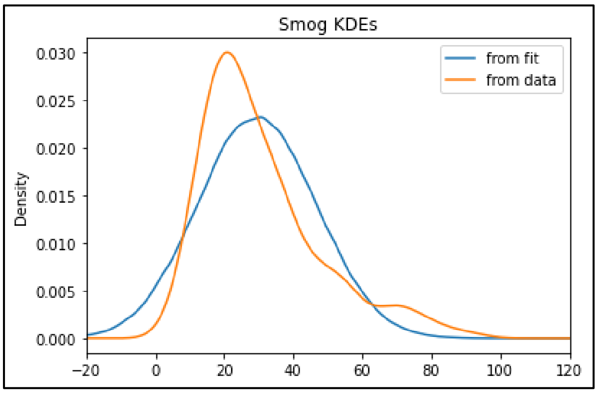
We received similar results for normal and student's t-distribution with wind speed input. We think that there are acceptable – densities have peaks in the same places for models and for data and tails are similar in both cases. Different situation is in models when we used models with wind-speed as parameter. Here better results are given by student's t-distribution but in both cases they aren't acceptable because peaks are in totally wrong places.

Bulwarowa station

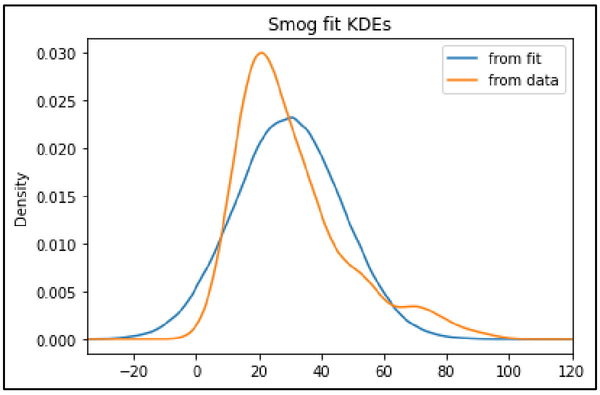
Normal distribution with wind speed modeled:



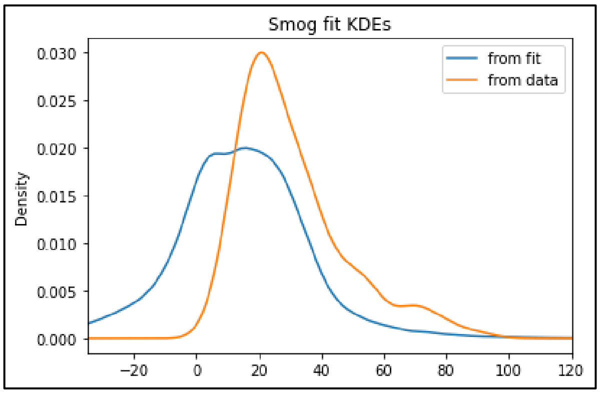
Normal distribution with wind speed input:



Student's t-distribution with wind-speed modeled:



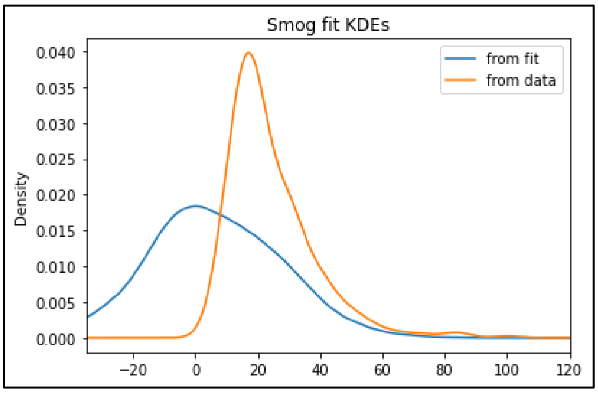
Student's t-distribution with wind speed input:



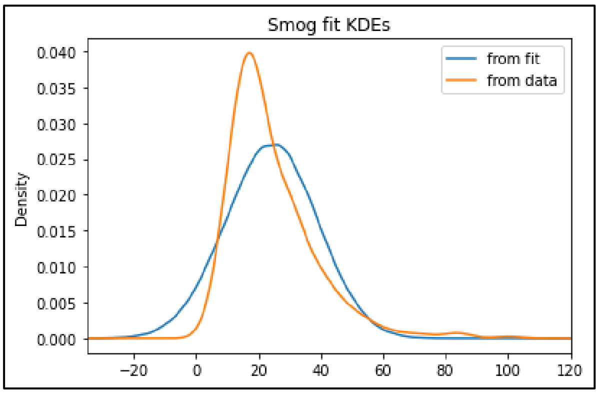
In Bulwarowa station situation is similar to Złoty Róg station. Data are concentrated and asymmetric while fittings have lower and quite shifted peak. We also received similar, acceptable results for models with wind speed input and wrong results for models with wind-speed as a parameter. But now results from models with wind speed as parameter are similar for both distributions.

Swoszowice station

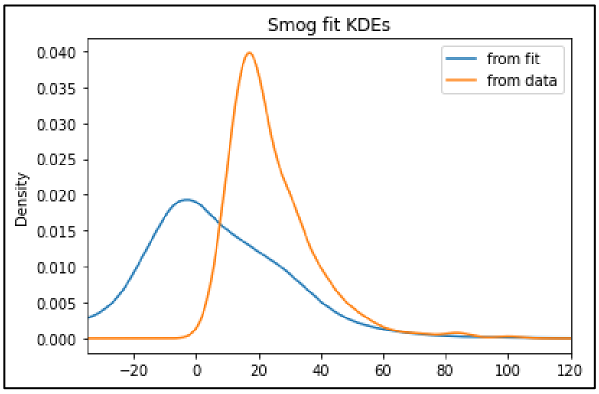
Normal distribution with wind speed modeled:



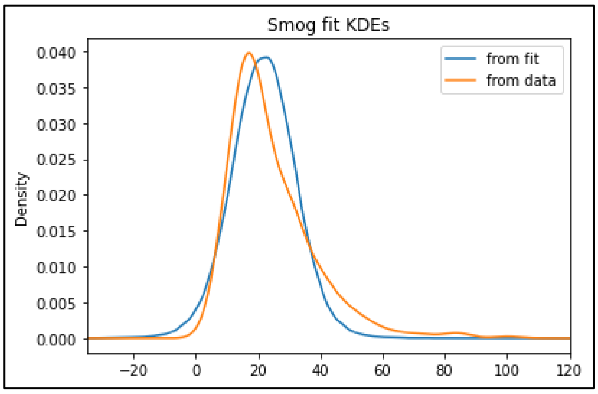
Normal distribution with wind speed input:



Student's distribution with wind-speed modeled:



Student's t-distribution with wind speed input:



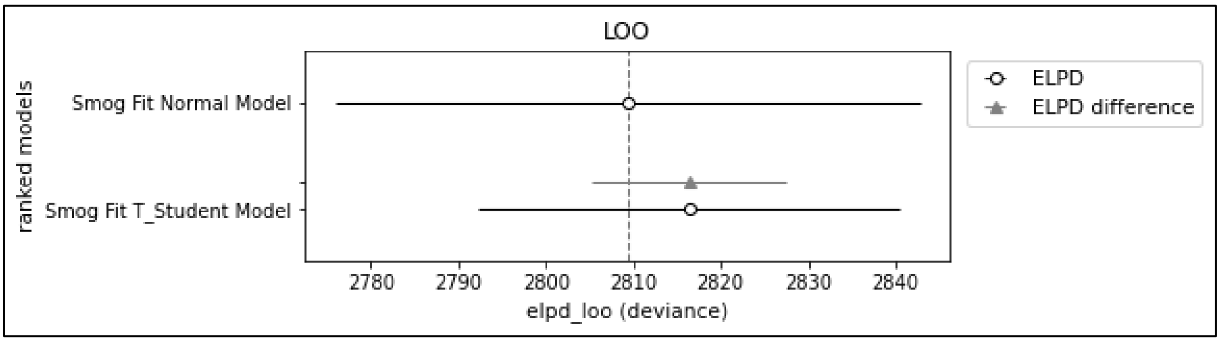
In Swoszoszowice station there is a different smog data distribution than in Bulwarowa or Złoty Róg stations and we also can see differences in models' performance. Data have more concentrated and symmetric values than in rest of stations. For models with wind speed as parameter we received results with very different densities than this from data. It isn't acceptable. In models with wind speed as input results are acceptable for both distributions but much better are for student's t-distribution where they are very similar to data and have peak with the same height as data.

We can see that in all cases models with wind speed as input perform much better than models with it as parameter so we decided to compare only second ones models. We can also conclude that preparing separate priors for every parameter and then connecting them in one posterior model may return wrong results even if separate results for every prior were acceptable. Better solution is to create whole model at a clip or use data not parameters if we have such possibility.

Model comparison

Model comparison for Złoty Róg Street

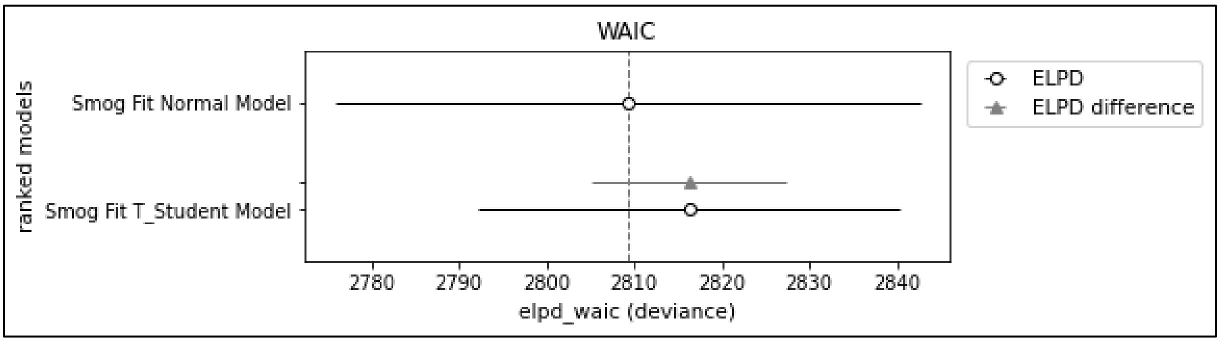
- **Result for PSIS_LOO**



	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
Smog Fit Normal Model	0	2809.316003	4.571082	0.000000	0.744819	33.482254	0.000000	False	deviance
Smog Fit T_Student Model	1	2816.349436	2.733748	7.033433	0.255181	24.049117	11.096589	False	deviance

Based on the ELPD values and weights, it can be seen that the "Smog Fit T_Student Model" has a higher ELPD and is assigned a higher weight compared to the "Smog Fit Normal Model." This suggests that the "Smog Fit T_Student Model" performs better in terms of the expected predictive performance. However, it's worth noting that the difference in ELPD between the two models is relatively small, and the standard errors be considered to assess the uncertainty in these estimates

- **Result for WAIC**

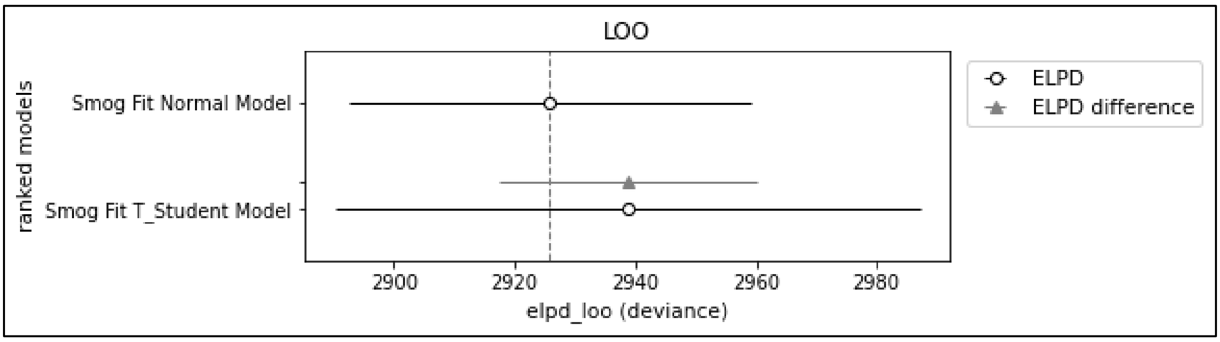


	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
Smog Fit T_Student Model	0	2752.221707	12.165902	0.000000	0.681015	56.570778	0.000000	False	deviance
Smog Fit Normal Model	1	2795.801565	6.412198	43.579859	0.318985	46.950066	30.771421	True	deviance

Based on the provided information the WAIC data suggests that the "Smog Fit T_Student Model" performs better than the "Smog Fit Normal Model" based on the estimated log pointwise predictive density. The T_Student Model has a higher weight and lower standard error compared to the Normal Model. However, it's important to note that the Normal Model has a significant difference in estimated log pointwise predictive density compared to the best-performing model. Additionally, there is a warning associated with the Normal Model, indicating a potential issue or concern.

Model comparison for Bulwarowa Street

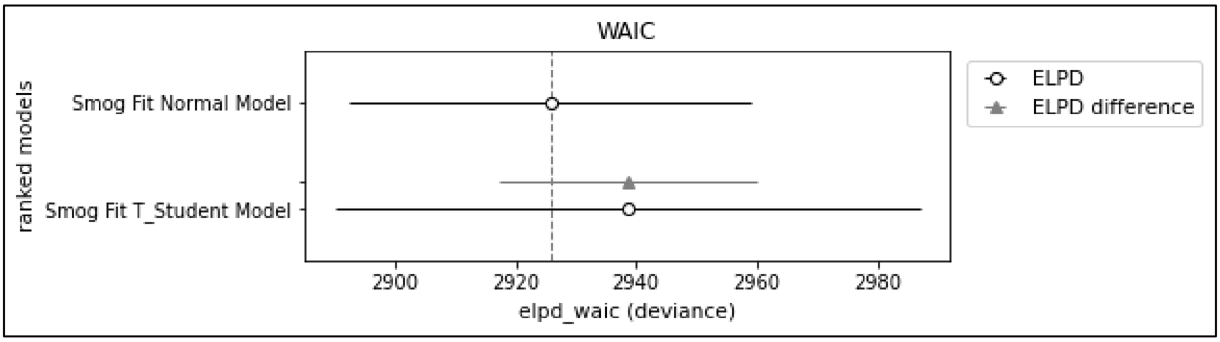
- **Result for PSIS_LOO**



	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
Smog Fit T_Student Model	0	2752.450037	12.280067	0.00000	0.680646	56.579885	0.000000	False	deviance
Smog Fit Normal Model	1	2795.846298	6.434565	43.39626	0.319354	46.937896	30.765342	False	deviance

Based on this interpretation, it seems that the "Smog Fit T_Student Model" outperforms the "Smog Fit Normal Model" in terms of elpd_loo, with a difference of 12.280067. However, the "Smog Fit Normal Model" has a lower p_loo value, indicating it is less complex. Additionally, no warnings were raised during the evaluation, and the estimation is likely related to the deviance measure.

- **Result for WAIC**



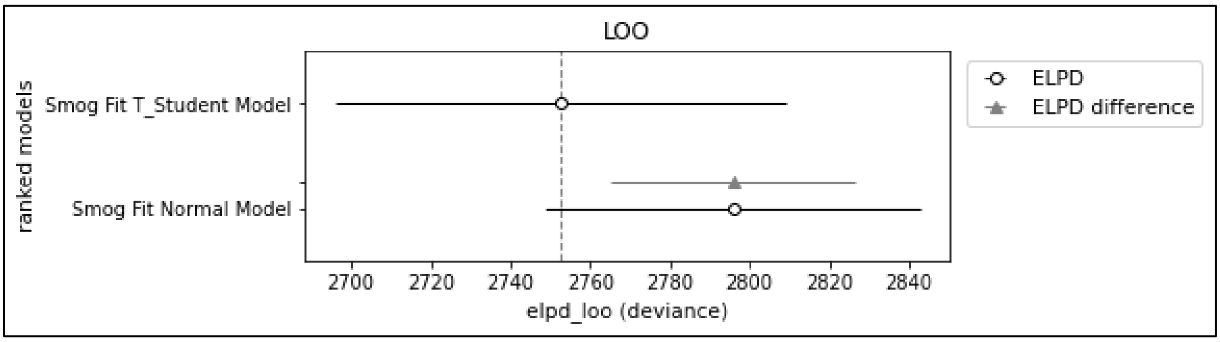
	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
Smog Fit T_Student Model	0	2752.221707	12.165902	0.000000	0.681015	56.570778	0.000000	False	deviance
Smog Fit Normal Model	1	2795.801565	6.412198	43.579859	0.318985	46.950066	30.771421	True	deviance

The "Smog Fit T_Student Model" has a slightly lower elpd_waic value compared to the "Smog Fit Normal Model," indicating that it may have a better fit to the data. However, the difference in elpd_waic is relatively small, suggesting that the two models perform similarly in terms of predictive performance.The weights assigned to each model indicate the probability of each model being the best performer, with the "Smog Fit T_Student Model" having a higher weight (0.681015) than the "Smog Fit Normal Model" (0.318985). This implies that there is more evidence supporting the "Smog Fit T_Student Model" as the better choice.

Overall, based on the provided data, the "Smog Fit T_Student Model" appears to be favored over the "Smog Fit Normal Model" in terms of the WAIC scores and weights assigned.

Model comparison for Swoszowice

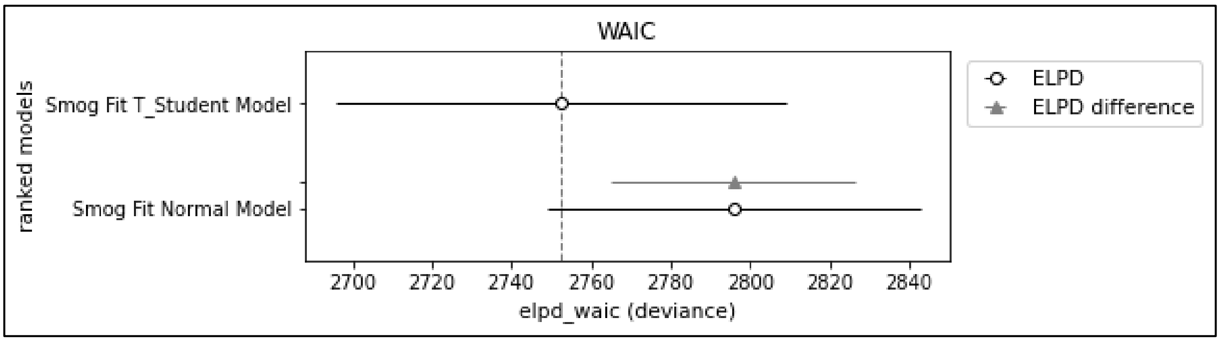
- **Result for PSIS_LOO**



	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
Smog Fit T_Student Model	0	2752.450037	12.280067	0.00000	0.680646	56.579885	0.000000	False	deviance
Smog Fit Normal Model	1	2795.846298	6.434565	43.39626	0.319354	46.937896	30.765342	False	deviance

Based on this interpretation, it seems that "Smog Fit T_Student Model" performs better than the "Smog Fit Normal Model" in terms of LOO statistics. It has a higher elpd_loo, lower p_loo, and a higher weight, indicating a higher likelihood of being the best model. However, the "Smog Fit Normal Model" still performs relatively well, with a positive elpd_diff compared to the best model.

- **Result for WAIC**



	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
Smog Fit T_Student Model	0	2752.221707	12.165902	0.000000	0.681015	56.570778	0.000000	False	deviance
Smog Fit Normal Model	1	2795.801565	6.412198	43.579859	0.318985	46.950066	30.771421	True	deviance

To sum up, the "Smog Fit T_Student Model" has a lower elpd_waic and a higher number of effective parameters (p_waic) compared to the "Smog Fit Normal Model." However, the "Smog Fit Normal Model" has a higher elpd, indicating better out-of-sample prediction accuracy. It also has a warning, which suggests potential issues with the model. The weight assigned to each model indicates the relative importance or confidence in their performance.