M. Fraszczyk, K. Kaczmarek-Majer, O. Hryniewicz, K. Skotak and A. Degórska, "Expert-in-the-loop Stepwise Regression and its Application in Air Pollution Modeling," 2022 IEEE 11th International Conference on Intelligent Systems (IS), Warsaw, Poland, 2022, pp. 1-7, doi: 10.1109/IS57118.2022.10019609

Expert-in-the-loop Stepwise Regression and its Application in Air Pollution Modeling

Miłosz Fraszczyk
Institute of Environmental Protection
National Research Institute,
Warsaw, Poland
fraszczyk.milosz@gmail.com

Katarzyna Kaczmarek-Majer
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
k.kaczmarek@ibspan.waw.pl
0000-0003-0422-9366

Olgierd Hryniewicz
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
hryniewi@ibspan.waw.pl
0000-0001-9877-508X

Krzysztof Skotak
Institute of Environmental Protection
National Research Institute,
Warsaw, Poland
krzysztof.skotak@ios.edu.pl

Anna Degórska
Institute of Environmental Protection
National Research Institute,
Warsaw, Poland
anna.degorska@ios.edu.pl

***Abstract*—In this work, we provide a statistical procedure to integrate expert preferences towards explanatory variables in stepwise forward regression. The proposed method builds on the traditional stepwise linear regression and goal programming. The procedure is validated experimentally for real-life data from various sources aiming at predicting air pollution. The practical goal is to predict the annual concentrations of two health-related air pollutants, namely PM10 (Particulate Matter that is 10 micrometers or less in diameter) and NO2 (Nitrogen Dioxide). The main finding from this work is that inclusion of expert knowledge leads to more robust and accurate predictive models. Considering the limited size of data from air pollution monitoring stations, additional expert knowledge enabled to select most meaningful explanatory variables, and as the consequence the statistical inference lead to the improved predictions. The main contribution of this work is the proposed simple but solid expert-in-the-loop stepwise forward linear regression method allowing to include expert preferences. Experiments confirm that the proposed procedure is not only more interpretable but also delivers more accurate predictions for the considered air pollutants concentrations.**

***Index Terms*—stepwise regression, linear programming, expert-in-the-loop, land-use regression, goal programming, intelligent data analysis, air pollution modeling**

I. Introduction

Recent study has shown that different statistical algorithms perform similarly when modelling annual average air pollution concentrations using a large number of training sites [1]. At the same time, one of the problems reported was missing variables resulting in limited size of the sample used for this supervised learning task. To alleviate this problem, we present a method that integrates knowledge of domain experts into the statistical inference.

In this work, we focus on the linear stepwise regression due to its simplicity. Contrary to the black-box algorithms, regression models are usually easier to interpret, both in terms of included predictors and the magnitude and direction of effects. As a starting point, we consider the supervised stepwise linear regression. Such class of models was applied previously for the considered air pollution monitoring problem, see e.g., de Hoogh et al. [2], [3]. However, we observed that if this procedure is performed on a relatively small samples, the resulting models might be overfitted. Therefore, the main contribution of this work is the proposed approach that enables to improve the training process with expert knowledge. In particular, we collect preferences expressed by the domain experts towards the explanatory variables to be included in the predictive modeling. Next, we run the goal programming to find the best balance between the quality of models (measured for example with the adjusted R2 of the training data) and the expert preference towards a variable.

The paper is organized as follows. In the second section, we present the motivating example for this work that is predicting air pollution concentrations to be considered for the epidemiological models. In the third section, we present the proposed statistical procedure that integrates expert knowledge into the predictive models. The main results of this research are presented in the fourth section of the paper where we demonstrate selected properties and a summary of models performance. The paper is concluded in its last section.

II. Motivating example: prediction of the air pollutants' annual concentrations

Air pollution affects various health issues all over the world. For example, in [4], the authors quantify this healthy burden for the area of Warsaw. A recent review [5] concludes that the observed increased air pollution levels are linked to increased general and respiratory disease mortality rates, higher prevalence of respiratory diseases, including asthma, lung cancer, etc. Characteristics of these links between air pollution and health are still subject to ongoing research. For example, the impact of air pollution on the developing brain is investigated within a recent NeuroSmog project [6] with focus on the area of southern Poland (see Figure 1). Among various air pollutants, PM10 (Particulate Matter that is 10 micrometers or less in diameter), PM2.5 and NO2 (Nitrogen Dioxide) are regarded to have an important impact on health aspects.

Although various strategies are implemented for mitigating emissions, see e.g., [7] concentrated on the area of Warsaw, according to the World Bank Group over half of the 50 most polluted cities in the European Union are located in Poland.

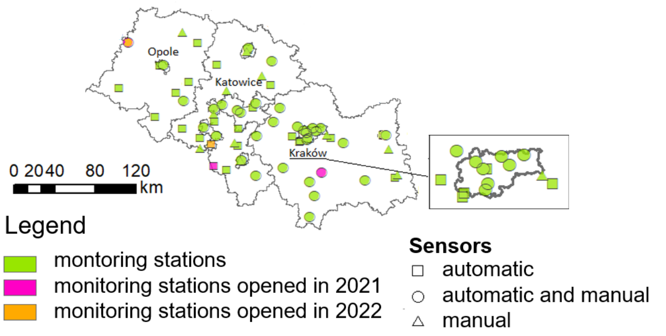


Figure 1. Visualisation of air pollution monitoring stations in the considered study area of the three voivodeship in the southern Poland (Silesian voivodeship, Opole voivodeship and Lesser Poland voivodeship).

The possibilities to collect data related to air pollutants from sensors have grown significantly in the recent years, see e.g., [8]. Nonetheless, epidemiologists often need air pollution estimates for historic years, and such sensor data are often not available for longer history. Furthermore, it also needs to be noted that various works on the physical models of air pollutants' distribution have been published, e.g., EMEP4PL [9]. However, atmospheric dispersion modelling often becomes computationally complex, and thus, infeasible to be performed with very fine spatial resolution. Also, in the considered applied scenario, only the annual estimates of the air pollution are needed.

III. The proposed method: Expert-in-the-loop stepwise linear regression

Let us consider the following multivariate linear regression:

$$y = X\beta + \epsilon \tag{1}$$

where y is an $m \times 1$ vector consisting of air pollutant concentrations, each of which corresponds to one monitoring station and the annual mean of the air pollutant. Measured concentrations of PM10 or NO2 are used as the response vectors, X is a $m \times s$ covariate matrix of the s explanatory variables (regression coefficients) related to the land-use, roads, emissions, etc., and ϵ is the vector of m errors which is assumed to be multivariate normal with zero mean. For more details related to linear regression models and statistical inference, see e.g., the seminal book of Hastie et al. [10].

Proper selection of subset of explanatory variables from X may become a challenging task if the size of the training data is limited. Thus, we extend the stepwise forward linear regression technique applied previously in this domain by de Hoogh et al. [2], [3]. We denote this model as **SLR**. In [2], a univariate linear regression model was run for each potential predictor to choose the model with the highest adjusted R2 as the starting point. Additional significant predictor variables were allowed to enter the model if they added to the adjusted R2 of the previous model step, and only if they adhered to the plausible direction of effect. Nonetheless, as we observe, if **SLR** procedure is performed on a relatively small samples, the resulting models might be overfitted.

Firstly, we propose to modify the procedure and to repeatedly check whether variables enter the models. We denote this approach as **ADV_SLR**. We consider again as candidates for variables those ones that have been already rejected in the previous loop, except for the predictors which are included in a temporary best model.

Secondly, as we observe, there are situations for both aforementioned methods when variables are similarly good candidates to enter the model, and the selection is based purely on adj R2 that might be affected by outliers. Thus, we introduce **Expert-in-the-loop Stepwise Linear Regression (EXP_SLR)** method and we add variables to model using the goal programming. The final model shall represent optimal performance and selection of variables preferred by expert.

The proposed **EXP_SLR** algorithm is a forward stepwise regression and Algorithm 1. details its steps. It continues to select the most adequate variable to enter the model using goal programming instead of the model with the highest metric. Goal programming considers model performance (e.g., R2) and expert preferences $pref\_exp$ towards each explanatory variable X. Those preferences need to be provided a priori, e.g., as a list of expert ratings. Additionally, user defines weight w_e expressing their preference towards expert knowledge over model evaluation metrics. Once we identify best candidate for next variable to enter the model, we validate whether the new model fulfils the required assumptions. In particular, p values of all variables are checked, sign of the coefficients is also validated (direction of effects need to meet the expectations of expert). Then, homoscedasticity of variance and normality of residuals are checked through analysis of residuals and inspection of quantile-quantile plots. Finally, the algorithm iterates until all candidates for variables are considered. The main steps of the goal programming are as follows:

1. Min-Max scale adjusted R2 from all models in current iteration
2. Min-Max scale preferences of expert
3. Calculating the distance dis_{R2} between scaled adjusted R2 and 1
4. Calculating the distance dis_{exp} between scaled preferences of expert and expedient value
5. Finding the minimal value for the equation:

$$w_{R2} * dis_{R2} + w_e * dis_{exp} \tag{2}$$

where w_e is the weight, and we calculate w_{R2}=1-w_e which is the weight for the objective evaluation metrics.

Algorithm 1. *Expert-in-the-loop Stepwise Linear Regression* **Input**: Y, data X, $pref\_exp$, α, w_e

Initialize: $temp\_rej$[], $selected$[], $temp\_best$

Output: $best_{model}$

```
1.1 $all\_candidates := \{pred_1, pred_2, \ldots, pred_n\}$
2.1 While TRUE:
      $selected$[] := ( set($all\_candidates$[]) - set($temp\_best$.variables) ) - set($temp\_rej$[])
  2.2.1 If $length$(selected[]) == 0:
          $best_{model}$ = $temp\_best$
          break
  2.2.2 For i in $selected$[]:
          $models$[] := LinearRegression(Y, $temp\_best$.variables,i)
  2.2.3 $model\_cand$ := GoalProgramming($models$[], $pref\_exp$, $w_e$)
        Unless univariate model:
  2.2.4 If (AdjR2($model\_cand$) - AdjR2($temp\_best$)) $< \alpha$ :
          $best_{model}$ = $temp\_best$
          break
  2.2.5 validate $model\_cand$:
    2.2.6.1 p values of predictors and direction of effects
    2.2.6.2 collinearity (VIF) - unless univariate model
    2.2.6.3 homoscedasticity of variance
    2.2.6.4 normality of residuals
  2.2.7 If $model\_cand$ validated:
          $temp\_best$ = $model\_cand$
          $temp\_rej$[].clear()
  2.2.8 Else if $model\_cand$ NOT validated:
          $temp\_rej$[].append($model\_cand$.variables)
```

$temp\_rej$[] - is a list of variables which are append in a situation, when searching model do not fulfil conditions. This list is cleared when any model is validated and after this, algorithm begin another loop.

$temp\_best$ - is a temporary the best validated model
$selected$[] - is a list of variables (candidates). There are some possibilities of containing candidates in it:

- all candidates if it is iteration finding univariate model
- all candidates without variables in $temp\_rej$[] if it still iterate to find properly univariate model.
- all candidates without variables in $temp\_best$ if it is new loop
- all candidates without variables in $temp\_best$ and without variables in $temp\_rej$[] if $temp\_best$ did not change after iteration

IV. Results

A. Explanatory variables

We illustrate the performance of the proposed **EXP_SLR** method for real-life data from various sources including air pollution monitoring stations, satellite images and sensors. The main groups of predictor variables considered in this study (and according to the study protocol, please see Table 3 in [6]) are as follows:
(i) traffic and residential emissions of air pollutants;
(ii) data about land-use from the Corine Land Cover database [11] including indicators such as, e.g., anthropogenic area, rural area, forest and wooded area, surface water, vegetation and agricultural area, undeveloped area;
(iii) road data, e.g., type of road (e.g., highway, expressway, main road etc.);
(iv) estimates from the atmospheric dispersion modelling following [9].

All explanatory variables are calculated for the following 7 grid cell sizes (from A to G):

- grid A: approx. 4km x 4km,
- grid B: approx. 2km x 2km,
- grid C: approx. 1km x 1km,
- grid D: approx. 500m x 500m,
- grid E: approx. 250m x 250m,
- grid F: approx. 125m x 125m
- grid G: approx. 62.5m x 62.5m.

Figure 2 illustrates the splitting mechanism.

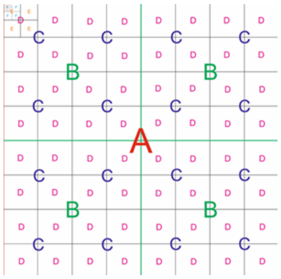


Figure 2. Illustration of region split into grid cell sizes

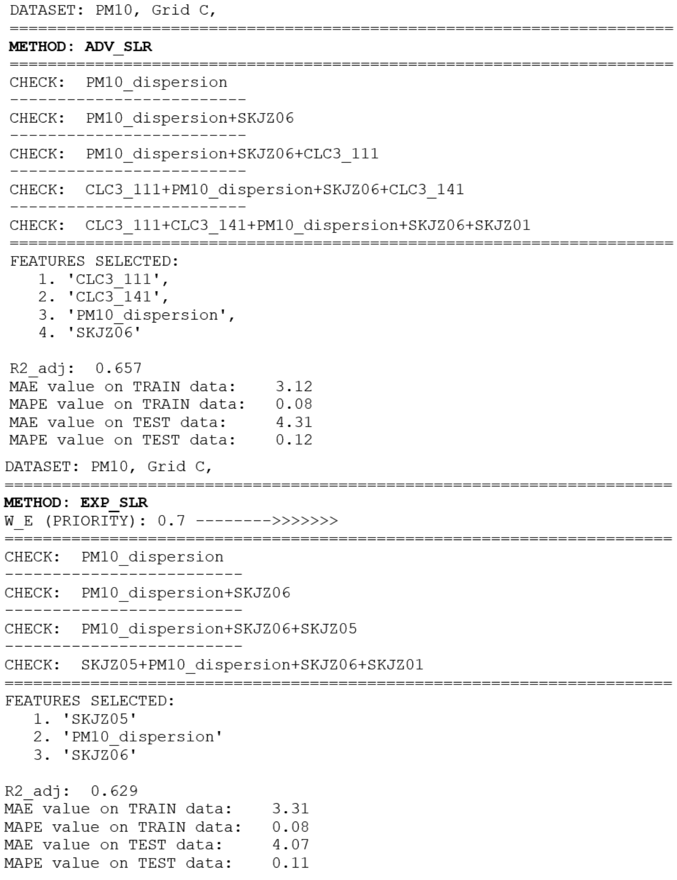
In experiments, we build stepwise regression models for various grid cell sizes independently to verify if the inclusion of expert knowledge is robust to the dataset spatial resolution (grid cell size).

B. Expert preferences

Data from domain expert were collected in the form of a three point scale. Options included low priority, medium priority and high priority. Expert evaluated his/her priority towards a variable to be included in predictive model. For example, all variables related to estimates from the atmospheric dispersion modelling were rated as high priority. Similarly was assessed main roads and wooded areas. On the contrary, some land-use variables such as e.g., surface water areas were considered as low priority. There were also several variables which expert marked as medium to high priority expressing his hesitation. We considered this rating as 2.5.

C. Illustrative example

We now provide an illustrative example of the step by step performance of the stepwise regression method based on crisp criteria such as e.g., R2. For example, let us consider data set for PM10 in grid C. Figure 3 shows the transcript from the inference with the baseline **ADV_SLR** method without expert preferences and from the regression with expert **EXP_SLR**.



```
 DATASET: PM10, Grid C,
 =======================================================================
 METHOD: ADV_SLR
 =======================================================================
 CHECK:  PM10_dispersion
 -------------------------
 CHECK:  PM10_dispersion+SKJZ06
 -------------------------
 CHECK:  PM10_dispersion+SKJZ06+CLC3_111
 -------------------------
 CHECK:  CLC3_111+PM10_dispersion+SKJZ06+CLC3_141
 -------------------------
 CHECK:  CLC3_111+CLC3_141+PM10_dispersion+SKJZ06+SKJZ01
 =======================================================================
 FEATURES SELECTED:
    1. 'CLC3_111',
    2. 'CLC3_141',
    3. 'PM10_dispersion',
    4. 'SKJZ06'

 R2_adj:  0.657
 MAE value on TRAIN data:    3.12
 MAPE value on TRAIN data:   0.08
 MAE value on TEST data:     4.31
 MAPE value on TEST data:    0.12
DATASET: PM10, Grid C,
=======================================================================
METHOD: EXP_SLR
W_E (PRIORITY): 0.7 -------->>>>>>>
=======================================================================
CHECK:  PM10_dispersion
-------------------------
CHECK:  PM10_dispersion+SKJZ06
-------------------------
CHECK:  PM10_dispersion+SKJZ06+SKJZ05
-------------------------
CHECK:  SKJZ05+PM10_dispersion+SKJZ06+SKJZ01
=======================================================================
FEATURES SELECTED:
   1. 'SKJZ05'
   2. 'PM10_dispersion'
   3. 'SKJZ06'

R2_adj:  0.629
MAE value on TRAIN data:    3.31
MAPE value on TRAIN data:   0.08
MAE value on TEST data:     4.07
MAPE value on TEST data:    0.11
```

Figure 3. Logs of the step by step performance of the **ADV_SLR** and **EXP_SLR** methods for an exemplary dataset PM10, grid C.

As observed in Fig. 3, for **ADV_SLR**, the PM10_dispersion was the first variable to enter the model. Next, SKJZ06 (local roads) had been chosen according to adj R2 and after validating the model, it also entered. Then, CLC3_111 (urban area) was included. Finally, CLC3_141 (green area) entered the model.

Next, let us consider that we assign expert weight w_e to 0.7. Let us see the step-by-step performance of the proposed **EXP_SLR** stepwise regression with expert-in-the-loop on the same dataset. Similarly, we start from PM10_dispersion. Then, the model adds SKJZ06 (local roads, priority high). However, instead of any land-use variable, the next check is related to another variable about roads that is SKJZ05 (main roads, priority high). We can also check that the variables related to the green areas and urban areas were assigned a medium priority by the expert. We see that the performance of **ADV_SLR** model is slightly higher for the training data, e.g., MAE of 3.12 for **ADV_SLR** and 3.31 for **EXP_SLR**). However, for the test data we observe the opposite relation, e.g., MAE of 4.31 for **ADV_SLR** and 4.07 for **EXP_SLR**).

D. Validation

The proposed method is designed to deliver more interpretable models by inclusion of variables that are most preferred. However, apart from the interpretability aspects, the accuracy of the proposed method needs to be investigated. In next experiments, we aim at discovering if inclusion of expert knowledge for the stepwise regression modeling influences the accuracy of predictions. Let us start from reviewing the R2 for the training set and checking whether inclusion of the expert knowledge influences the training phase. Table I shows R2 for the TRAIN set.

Results in Table I are averages from models built on 80% of the monitoring sites with the remaining 20% used for validation. To evaluate the performance of the proposed approach, we repeated 10 times the random selection of training and test sets. As observed, the highest mean R2 (averaged across all grid cell sizes) for NO2 amounts 0.79 and is observed for **SLR** and **ADV_SLR** methods. For PM10 it amounts to 0.59 and is achieved by the **ADV_SLR** method. The second highest R2 are usually obtained by **EXP_SLR** with w_e=0.5 and 0.6. The differences in R2 are overall small considering that these results are calculated for the training sets.

Next, to test the robustness and the stability of the procedure, we calculate the mean absolute error (MAE) and mean absolute percentage error (MAPE). Table II and Table III show these metrics for the TRAIN and TEST sets of PM10 and NO2, respectively.

An average of mean absolute error with PM10 and NO2 of **SLR** is almost always better on train data of each grid cell size (A, B, C ...). The opposite situation is observed on the TEST data. For example, for the TRAIN data, the average MAE is smallest for the **ADV_SLR** method and amounts to 3.43. At the same time, for the considered PM10, the mean MAE for the TEST set amounts to 5.26 and is achieved by the **EXP_SLR** method with weights of 0.8 and 0.9. For the TEST set, MAPE is also smaller for the **EXP_SLR** method than for the **SLR**. Similarly, we observe that standard deviation of errors is smaller for the method with expert preferences.

Table I

ADJUSTED R2 ON TRAIN DATA (AVERAGED FROM K=10 ITERATIONS) AT VARIOUS GRID CELL SIZES WHEN MODELING NO2 AND PM10 CONCENTRATIONS WITH STEPWISE LINEAR REGRESSION **SLR**, STEPWISE LINEAR REGRESSION WITH MORE REPETITIONS TO INCLUDE NEW VARIABLES IN MODEL **ADV_SLR** AND THE PROPOSED STEPWISE LINEAR REGRESSION WITH EXPERT PREFERENCES **EXP_SLR** WITH WEIGHTS w_e FROM $\{0.5., 0.6, 0.7, 0.8, 0.9\}$. TRAIN ARE THE MEAN VALUES ACROSS ALL GRID CELL SIZED CALCULATED FOR THE TRAIN SET.

NO2 Adj R2									
Method		A	B	C	D	E	F	G	TRAIN
SLR		0.88	0.84	0.82	0.74	0.75	0.76	0.75	0.79
ADV_SLR		0.88	0.84	0.82	0.74	0.75	0.76	0.75	0.79
EXP_SLR	w_e = 0.5	0.87	0.85	0.80	0.73	0.74	0.76	0.75	0.78
EXP_SLR	w_e = 0.6	0.86	0.85	0.79	0.73	0.74	0.76	0.75	0.78
EXP_SLR	w_e = 0.7	0.83	0.82	0.76	0.73	0.73	0.75	0.75	0.77
EXP_SLR	w_e = 0.8	0.77	0.71	0.74	0.73	0.67	0.75	0.74	0.73
EXP_SLR	w_e = 0.9	0.77	0.71	0.74	0.73	0.67	0.75	0.74	0.73
PM10 Adj R2									
Method		A	B	C	D	E	F	G	TRAIN
SLR		0.56	0.59	0.61	0.59	0.59	0.58	0.55	0.58
ADV_SLR		0.56	0.59	0.64	0.59	0.59	0.59	0.56	0.59
EXP_SLR	w_e = 0.5	0.54	0.59	0.64	0.59	0.58	0.58	0.56	0.58
EXP_SLR	w_e = 0.6	0.52	0.59	0.64	0.59	0.58	0.57	0.55	0.58
EXP_SLR	w_e = 0.7	0.52	0.57	0.63	0.58	0.56	0.56	0.54	0.56
EXP_SLR	w_e = 0.8	0.51	0.55	0.60	0.58	0.56	0.56	0.54	0.56
EXP_SLR	w_e = 0.9	0.51	0.55	0.60	0.58	0.56	0.56	0.54	0.56
NO2 std dev of Adj R2									
Method		A	B	C	D	E	F	G	TRAIN
SLR		0.02	0.04	0.08	0.07	0.07	0.07	0.05	0.06
ADV_SLR		0.02	0.05	0.08	0.07	0.07	0.07	0.05	0.06
EXP_SLR	w_e = 0.5	0.03	0.04	0.08	0.08	0.06	0.07	0.06	0.06
EXP_SLR	w_e = 0.6	0.03	0.04	0.08	0.08	0.06	0.07	0.06	0.06
EXP_SLR	w_e = 0.7	0.07	0.06	0.08	0.08	0.06	0.07	0.06	0.07
EXP_SLR	w_e = 0.8	0.08	0.08	0.07	0.08	0.07	0.07	0.07	0.07
EXP_SLR	w_e = 0.9	0.08	0.09	0.07	0.08	0.07	0.07	0.07	0.07
PM10 std dev of Adj R2									
Method		A	B	C	D	E	F	G	TRAIN
SLR		0.26	0.19	0.13	0.16	0.20	0.25	0.23	0.20
ADV_SLR		0.26	0.19	0.13	0.16	0.20	0.26	0.23	0.20
EXP_SLR	w_e = 0.5	0.25	0.19	0.13	0.16	0.21	0.25	0.23	0.20
EXP_SLR	w_e = 0.6	0.23	0.19	0.13	0.16	0.21	0.25	0.22	0.20
EXP_SLR	w_e = 0.7	0.23	0.18	0.14	0.18	0.21	0.26	0.25	0.21
EXP_SLR	w_e = 0.8	0.23	0.18	0.13	0.18	0.21	0.26	0.25	0.21
EXP_SLR	w_e = 0.9	0.22	0.18	0.13	0.18	0.21	0.26	0.25	0.20

Finally, the differences between MAE of mean train and test data (the last two columns) are smaller with EXP_SLR methods with various weights than SLR method. It might be related to the fact that this method allows better generalisation. Exactly the same situation is apparent with results of mean absolute percentage error as in MAE.

Experimental results gathered in Tables II and III show also that for NO2 there are more models overtrained than for PM10 due to low amount of training data. for some meshes it is better to use the function with an expert, because though adj R2 may be slightly worse for some particular grid cell size, MAE and MAPE for the TEST set return equal or better.

The proposed method requires to set up prior weight towards expert knowledge. Figure 4 show results of simulations that aimed to show the influence of prior weights on the outputs.

Plots gathered in Figure 4 illustrate with simulations the relation between the prior parameters w_e and the expert preference towards the candidate variable on the objective function. Each line corresponds to one option regarding expert preferences (preference 1 is for low, 2 for medium, 2.5 for medium/high and 3 for high). Lines show the ranges from the minimum (i.e. the best) to maximum (i.e. worst) value for the equation $w_{R2} * dis_{R2} + w_e * dis_{exp}$, what the feature of the given priority can achieve. Distance dis_{R2} is our $absolute[(Min-maxscaledR2adjusted)-1]$ and dis_{exp} is $absolute[(Min-maxscaledpreferences)-expedient\_value]$. We observe various slopes of lines depending on what are the values of the initial w_e. The higher weight w_e we confer, the smaller validity of dis_R2 is in the equation. The difference between distances and positions of each line is defined by dis_{exp}.

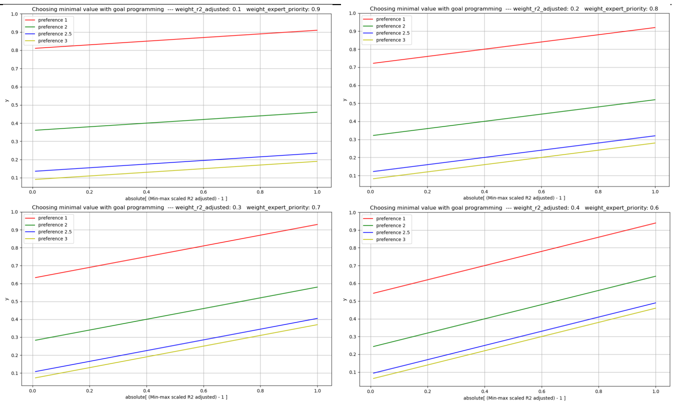


Figure 4. Simulated impact of w_e on the objective function and the expert preference towards the candidate variable. Four scenarios are considered having w_e=0.6, 0.7, 0.8, 0.9, respectively.

Moreover, assigning high w_e can lead to the situation in which less important preferences will not be considered during stepwise linear regression. Therefore, searching for the best combination of weights and preferences is crucial in order to find better models than **SLR** is able to return.

V. CONCLUSION AND FURTHER WORK

Collecting data from various sources may not always be effective, in particular for historic years. For this reason, domain expert knowledge might be of help to improve the processing of numerical datasets and statistical learning. Therefore, various expert-in-the-loop approaches are gaining attention in recent years. However, expert knowledge and statistical inference need to be carefully integrated. In this work, we introduced Expert-in-the-loop Stepwise Linear Regression. Feature selection was enhanced with expert preferences and goal programming. We presented a use case in predicting annual concentrations of two health-related air pollutants.

Future directions to improve the proposed approach include extensions to the semi-supervised learning variant, e.g., with data from sensors. Another idea is to improve the communication with experts and modeling of their expert preferences. Expert knowledge could be represented in a more advanced way, e.g., with the use of fuzzy sets. Finally, inclusion of other important health-related air pollutants such as PM2.5, combination of data of different spatial resolution (grid cell sized), and handling of spatial autocorrelations with statistical inference approach, see e.g., [12] remain open for future work.

Table II

MAE and MAPE comparison of supervised stepwise linear regression **SLR**, advanced SLR with repeated procedure to include variables **ADV_SLR** and stepwise linear regression with expert preferences **EXP_SLR**. Weights w_e : 0.5., 0.6, 0.7, 0.8, 0.9 are considered. Experiments are averaged for prediction of PM10 concentrations in 2018 averaged across monitoring sites of the NeuroSmog domain. TRAIN and TEST are the mean values across all grid cell sized calculated for the train and test sets, respectively.

PM10 Mean of MAE																	
Method		A	A_test	B	B_test	C	C_test	D	D_test	E	E_test	F	F_test	G	G_test	TRAIN	TEST
SLR		3.52	5.29	3.42	5.08	3.42	5.18	3.58	6.03	3.33	6.32	3.36	4.62	3.53	5.80	3.45	5.47
ADV_SLR		3.52	5.29	3.42	5.08	3.34	4.96	3.58	6.03	3.33	6.32	3.33	4.61	3.49	5.80	3.43	5.44
EXP_SLR	w_e = 0.5	3.63	5.39	3.42	5.10	3.34	5.10	3.62	5.94	3.42	6.46	3.42	4.74	3.48	5.79	3.48	5.50
EXP_SLR	w_e = 0.6	3.74	5.36	3.47	5.12	3.34	5.10	3.62	5.94	3.51	6.54	3.47	4.74	3.53	5.78	3.52	5.51
EXP_SLR	w_e = 0.7	3.74	5.28	3.60	5.13	3.41	5.06	3.69	6.03	3.62	6.42	3.57	4.71	3.63	4.70	3.61	5.33
EXP_SLR	w_e = 0.8	3.76	5.23	3.71	5.10	3.56	4.65	3.69	6.03	3.62	6.42	3.57	4.71	3.63	4.70	3.65	5.26
EXP_SLR	w_e = 0.9	3.80	5.22	3.73	5.06	3.56	4.65	3.69	6.03	3.62	6.42	3.57	4.71	3.63	4.70	3.66	5.26
PM10 Mean of MAPE																	
Method		A	A_test	B	B_test	C	C_test	D	D_test	E	E_test	F	F_test	G	G_test	TRAIN	TEST
SLR		0.09	0.14	0.09	0.13	0.09	0.14	0.09	0.17	0.09	0.18	0.09	0.12	0.09	0.17	0.09	0.15
ADV_SLR		0.09	0.14	0.09	0.13	0.09	0.13	0.09	0.17	0.09	0.18	0.09	0.12	0.09	0.17	0.09	0.15
EXP_SLR	w_e = 0.5	0.10	0.14	0.09	0.13	0.09	0.14	0.09	0.17	0.09	0.18	0.09	0.12	0.09	0.17	0.09	0.15
EXP_SLR	w_e = 0.6	0.10	0.14	0.09	0.13	0.09	0.14	0.09	0.17	0.09	0.19	0.09	0.12	0.09	0.17	0.09	0.15
EXP_SLR	w_e = 0.7	0.10	0.14	0.09	0.13	0.09	0.14	0.10	0.17	0.09	0.18	0.09	0.12	0.10	0.12	0.09	0.14
EXP_SLR	w_e = 0.8	0.10	0.14	0.10	0.13	0.09	0.12	0.10	0.17	0.09	0.18	0.09	0.12	0.10	0.12	0.10	0.14
EXP_SLR	w_e = 0.9	0.10	0.14	0.10	0.13	0.09	0.12	0.10	0.17	0.09	0.18	0.09	0.12	0.10	0.12	0.10	0.14
PM10 Standard deviation of MAE																	
Method		A	A_test	B	B_test	C	C_test	D	D_test	E	E_test	F	F_test	G	G_test	TRAIN	TEST
SLR		0.88	1.55	0.80	1.32	0.40	1.56	0.61	4.18	0.82	4.41	0.92	1.45	0.72	3.39	0.74	2.55
ADV_SLR		0.88	1.55	0.80	1.32	0.40	1.28	0.61	4.18	0.82	4.41	0.94	1.45	0.72	3.39	0.74	2.51
EXP_SLR	w_e = 0.5	0.84	1.49	0.80	1.32	0.40	1.29	0.58	4.23	0.87	4.21	0.90	1.40	0.73	3.39	0.73	2.48
EXP_SLR	w_e = 0.6	0.76	1.50	0.78	1.30	0.40	1.29	0.58	4.23	0.86	4.15	0.90	1.40	0.70	3.40	0.71	2.47
EXP_SLR	w_e = 0.7	0.76	1.56	0.71	1.22	0.42	1.23	0.65	4.71	0.89	4.15	0.90	1.38	0.87	1.34	0.74	2.23
EXP_SLR	w_e = 0.8	0.74	1.56	0.71	1.08	0.32	1.28	0.65	4.71	0.89	4.15	0.90	1.38	0.87	1.34	0.73	2.21
EXP_SLR	w_e = 0.9	0.72	1.57	0.69	1.10	0.32	1.28	0.65	4.71	0.89	4.15	0.90	1.38	0.87	1.34	0.72	2.22
PM10 Standard deviation of MAPE																	
Method		A	A_test	B	B_test	C	C_test	D	D_test	E	E_test	F	F_test	G	G_test	TRAIN	TEST
SLR		0.03	0.04	0.02	0.03	0.01	0.05	0.02	0.13	0.02	0.15	0.03	0.04	0.02	0.13	0.02	0.08
ADV_SLR		0.03	0.04	0.02	0.03	0.01	0.04	0.02	0.13	0.02	0.15	0.03	0.04	0.02	0.13	0.02	0.08
EXP_SLR	w_e = 0.5	0.03	0.04	0.02	0.03	0.01	0.04	0.02	0.14	0.02	0.15	0.03	0.04	0.02	0.13	0.02	0.08
EXP_SLR	w_e = 0.6	0.02	0.04	0.02	0.03	0.01	0.04	0.02	0.14	0.02	0.15	0.03	0.04	0.02	0.13	0.02	0.08
EXP_SLR	w_e = 0.7	0.02	0.04	0.02	0.03	0.01	0.04	0.02	0.15	0.02	0.15	0.03	0.04	0.03	0.04	0.02	0.07
EXP_SLR	w_e = 0.8	0.02	0.04	0.02	0.03	0.01	0.03	0.02	0.15	0.02	0.15	0.03	0.04	0.03	0.04	0.02	0.07
EXP_SLR	w_e = 0.9	0.02	0.04	0.02	0.03	0.01	0.03	0.02	0.15	0.02	0.15	0.03	0.04	0.03	0.04	0.02	0.07

Acknowledgment

Miłosz Fraszczyk, Katarzyna Kaczmarek-Majer, Anna Degórska and Krzysztof Skotak received funding from the Foundation for Polish Science through the "NeuroSmog: Determining the impact of air pollution on the developing brain" (Nr. POIR.04.04.00-1763/18-00) project which is implemented as part of the TEAM-NET programme of the Foundation for Polish Science, co-financed from EU resources, obtained from the European Regional Development Fund under the Smart Growth Operational Programme. The authors thank Jakub Bratkowski and Julia Soloch for their support in data preparation.

References

[1] J. Chen, K. de Hoogh, J. Gulliver, B. Hoffmann, O. Hertel, M. Ketzel, M. Bauwelinck, A. van Donkelaar, U. A. Hvidtfeldt, K. Katsouyanni, N. A. Janssen, R. V. Martin, E. Samoli, P. E. Schwartz, M. Stafoggia, T. Bellander, M. Strak, K. Wolf, D. Vienneau, R. Vermeulen, B. Brunekreef, and G. Hoek, "A comparison of linear regression, regularization, and machine learning algorithms to develop europe-wide spatial models of fine particles and nitrogen dioxide," *Environment International*, vol. 130, p. 104934, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160412019304404

[2] de Hoogh K, W. M, A. M, B. C, B. R, B. M, C. G, and C. M, "Development of land use regression models for particle composition in twenty study areas in europe," *Environ Sci Technol*, vol. 47, pp. 5778–86, 2013.

[3] K. de Hoogh, J. Chen, J. Gulliver, B. Hoffmann, O. Hertel, M. Ketzel, M. Bauwelinck, A. van Donkelaar, U. A. Hvidtfeldt, K. Katsouyanni, J. Klompmaker, R. V. Martin, E. Samoli, P. E. Schwartz, M. Stafoggia, T. Bellander, M. Strak, K. Wolf, D. Vienneau, B. Brunekreef, and G. Hoek, "Spatial pm2.5, no2, o3 and bc models for western europe – evaluation of spatiotemporal stability," *Environment International*, vol. 120, pp. 81–92, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160412018309759

[4] P. Holnicki, M. Tainio, A. Kałuszko, and Z. Nahorski, "Burden of mortality and disease attributable to multiple air pollutants in warsaw, poland," *International Journal of Environmental Research and Public Health*, vol. 14, no. 11, 2017. [Online]. Available: https://www.mdpi.com/1660-4601/14/11/1359

[5] W. Nazar and M. Niedoszytko, "Air pollution in poland: A 2022 narrative review with focus on respiratory diseases," *International Journal of Environmental Research and Public Health*, vol. 19, no. 2, 2022. [Online]. Available: https://www.mdpi.com/1660-4601/19/2/895

[6] I. Markevych, N. Orlov, J. Grellier, K. Kaczmarek-Majer, M. Lipowska, K. Sitnik-Warchulska, Y. Mysak, C. Baumbach, M. Wierzba-Łukaszyk, M. H. Soomro, M. Compa, B. Izydorczyk, K. Skotak, A. Degórska,

Table III

MAE AND MAPE COMPARISON OF SUPERVISED STEPWISE LINEAR REGRESSION **SLR**, ADVANCED SLR WITH REPEATED PROCEDURE TO INCLUDE VARIABLES **ADV_SLR** AND STEPWISE LINEAR REGRESSION WITH EXPERT PREFERENCES **EXP_SLR**. WEIGHTS $w\_e : 0.5., 0.6, 0.7, 0.8, 0.9$ ARE CONSIDERED. EXPERIMENTS ARE AVERAGED FOR PREDICTION OF NO2 CONCENTRATIONS IN 2018 AVERAGED ACROSS MONITORING SITES OF NEUROSMOG DOMAIN. TRAIN AND TEST ARE THE MEAN VALUES ACROSS ALL GRID CELL SIZED CALCULATED FOR THE TRAIN AND TEST SETS, RESPECTIVELY.

NO2 Mean of MAE																	
Method		A	A_test	B	B_test	C	C_test	D	D_test	E	E_test	F	F_test	G	G_test	TRAIN	TEST
SLR		2.81	7.98	3.21	6.67	3.29	6.48	4.33	6.86	4.37	7.88	4.00	6.50	4.25	6.73	3.75	7.02
ADV_SLR		2.81	7.98	3.18	6.64	3.29	6.48	4.33	6.86	4.37	7.88	4.00	6.50	4.25	6.73	3.75	7.01
EXP_SLR	w_e = 0.5	2.93	8.09	3.17	6.97	3.58	5.52	4.44	6.78	4.44	7.79	4.09	6.10	4.20	6.44	3.83	6.81
EXP_SLR	w_e = 0.6	3.00	8.09	3.17	6.97	3.65	5.53	4.44	6.78	4.44	7.79	4.09	6.10	4.20	6.44	3.85	6.81
EXP_SLR	w_e = 0.7	3.33	6.74	3.34	6.84	3.96	5.50	4.44	6.78	4.64	8.27	4.18	5.84	4.20	6.44	4.01	6.63
EXP_SLR	w_e = 0.8	3.63	6.40	3.94	7.35	4.12	5.55	4.44	6.78	5.13	7.97	4.21	5.83	4.28	6.15	4.25	6.58
EXP_SLR	w_e = 0.9	3.63	6.40	3.94	7.49	4.12	5.55	4.44	6.78	5.13	7.97	4.21	5.83	4.28	6.15	4.25	6.60
NO2 Mean of MAPE																	
Method		A	A_test	B	B_test	C	C_test	D	D_test	E	E_test	F	F_test	G	G_test	TRAIN	TEST
SLR		0.14	0.34	0.17	0.29	0.17	0.33	0.23	0.33	0.22	0.34	0.19	0.31	0.22	0.32	0.19	0.32
ADV_SLR		0.14	0.34	0.17	0.29	0.17	0.33	0.23	0.33	0.22	0.34	0.19	0.31	0.22	0.32	0.19	0.32
EXP_SLR	w_e = 0.5	0.15	0.33	0.17	0.29	0.19	0.27	0.23	0.33	0.22	0.33	0.21	0.29	0.22	0.30	0.20	0.31
EXP_SLR	w_e = 0.6	0.16	0.33	0.17	0.29	0.19	0.27	0.23	0.33	0.22	0.33	0.21	0.29	0.22	0.30	0.20	0.31
EXP_SLR	w_e = 0.7	0.17	0.31	0.18	0.29	0.20	0.27	0.23	0.33	0.23	0.34	0.22	0.27	0.22	0.30	0.21	0.30
EXP_SLR	w_e = 0.8	0.18	0.31	0.20	0.32	0.20	0.27	0.23	0.33	0.25	0.34	0.23	0.27	0.23	0.29	0.22	0.30
EXP_SLR	w_e = 0.9	0.18	0.31	0.20	0.32	0.20	0.27	0.23	0.33	0.25	0.34	0.23	0.27	0.23	0.29	0.22	0.30
NO2 Standard deviation of MAE																	
Method		A	A_test	B	B_test	C	C_test	D	D_test	E	E_test	F	F_test	G	G_test	TRAIN	TEST
SLR		0.56	1.97	0.40	2.67	0.84	2.74	0.62	1.41	0.78	2.38	0.44	2.03	0.73	1.16	0.62	2.05
ADV_SLR		0.56	1.97	0.45	2.58	0.84	2.74	0.62	1.41	0.78	2.38	0.44	2.03	0.73	1.16	0.63	2.04
EXP_SLR	w_e = 0.5	0.56	1.73	0.49	2.99	0.76	1.30	0.67	1.50	0.68	2.32	0.52	2.17	0.83	1.14	0.64	1.88
EXP_SLR	w_e = 0.6	0.48	1.73	0.49	2.99	0.78	1.32	0.67	1.50	0.68	2.32	0.52	2.17	0.83	1.14	0.64	1.88
EXP_SLR	w_e = 0.7	0.54	2.09	0.63	2.99	0.67	1.09	0.67	1.50	0.60	3.01	0.54	2.06	0.83	1.14	0.64	1.98
EXP_SLR	w_e = 0.8	0.74	1.88	0.77	2.40	0.50	1.08	0.67	1.50	0.98	2.90	0.48	2.05	0.89	1.04	0.72	1.84
EXP_SLR	w_e = 0.9	0.74	1.88	0.78	2.59	0.50	1.08	0.67	1.50	0.98	2.90	0.48	2.05	0.89	1.04	0.72	1.86
NO2 Standard deviation of MAPE																	
Method		A	A_test	B	B_test	C	C_test	D	D_test	E	E_test	F	F_test	G	G_test	TRAIN	TEST
SLR		0.02	0.08	0.03	0.14	0.05	0.19	0.04	0.10	0.04	0.08	0.03	0.09	0.04	0.08	0.04	0.11
ADV_SLR		0.02	0.08	0.03	0.13	0.05	0.19	0.04	0.10	0.04	0.08	0.03	0.09	0.04	0.08	0.04	0.11
EXP_SLR	w_e = 0.5	0.02	0.10	0.04	0.12	0.05	0.11	0.04	0.10	0.03	0.08	0.04	0.09	0.04	0.08	0.04	0.10
EXP_SLR	w_e = 0.6	0.02	0.10	0.04	0.12	0.05	0.12	0.04	0.10	0.03	0.08	0.04	0.09	0.04	0.08	0.04	0.10
EXP_SLR	w_e = 0.7	0.03	0.10	0.05	0.12	0.04	0.10	0.04	0.10	0.03	0.08	0.03	0.08	0.04	0.08	0.04	0.09
EXP_SLR	w_e = 0.8	0.03	0.11	0.05	0.13	0.03	0.09	0.04	0.10	0.04	0.09	0.03	0.08	0.05	0.08	0.04	0.10
EXP_SLR	w_e = 0.9	0.03	0.11	0.05	0.14	0.03	0.09	0.04	0.10	0.04	0.09	0.03	0.08	0.05	0.08	0.04	0.10

J. Bratkowski, B. Kossowski, A. Domagalik, and M. Szwed, "Neurosmog: Determining the impact of air pollution on the developing brain: project protocol," *Int J Environ Res Public Health*, 2022.

[7] P. Holnicki, A. Kałuszko, and Z. Nahorski, "Analysis of emission abatement scenario to improve urban air quality," *Archives of Environmental Protection*, 2021.

[8] M. Kusy, P. Kowalski, M. Szwagrzyk, and A. Konior, "Machine learning techniques for explaining air pollution prediction," in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2022.

[9] W. M, M. Kryza, and K. Wałaszek, "Emission projections and limit values of air pollution concentration - a case study using the emep4pl model international," *Journal of Environment and Pollution*, vol. 65, 2019.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *Linear Methods for Regression*. New York, NY: Springer New York, 2009, pp. 43–99.

[11] CLC, "Corine land cover https://land.copernicus.eu/," *Version 2020 20u1*, 2018.

[12] J. Horabik and Z. Nahorski, "Improving resolution of a spatial air pollution inventory with a statistical inference approach," *Climatic Change*, 2014.