KOMUNIKATY NAUKOWE

Krystyna KONCA-KĘDZIERSKA

Institute of Meteorology and Water Management
National Research Institute

APPLICATION OF PARTIAL REGRESSION METHODS TO LONG RANGE FORECASTS

Description of the problem

From the beginning of time, people have relied on the weather prediction – to plan for adequate shelter, to organize their journey and how to go about farming. Today, we seek reliable weather forecasts in almost every aspect of our lives – including planning vacations or deciding when to start harvest. Not to mention the role of forecasts in aerospace transit. Large deposits of data, together with modern numerical models, satisfy this need in terms of short range (few days) weather forecasts. Such forecasts can be found on the Weather Forecast Service of Institute of Meteorology and Water Managemenet – National Research Institute (IMGW-PIB; 5 days[1]), MojaPogoda (MyWeather) – a weather forecast service of the MeteoGroup (4 days[2]) and the Numerical Weather Forecast Interdisciplinary Centre for Mathematical and Computational Modelling UW (ICM-UW; 3 days[3]). For longer periods, a simple integration of the equations of motion is no longer an appropriate approach, as the laws of deterministic chaos are no longer negligible (Silingo, Palmer 2011). The problem of long-range forecasts is defined as the estimation of the nature of the weather over a long period of time, and for this, statistical methods have to be applied. Many institutions are currently working on creating an accurate model for long-range weather forecasts. In their efforts, they rely on various probabilistic and deterministic

[1] http://www.pogodynka.pl/
[2] http://www.mojapogoda.com/pogoda.html
[3] http://www.meteo.pl/

ensemble-based methods and numerous cutting-edge numerical and statistical tools. The World Meteorological Organization (WMO[4]) has published the list of 13 institutions chosen for generating seasonal forecasts (WMO_gpc). The European Centre for Medium-Range Weather Forecasts (ECMWF[5]) publishes global probabilistic forecasts of the anomaly levels for monthly mean temperature at 2 m, mean atmospheric pressure at sea level, the monthly sum of the precipitation and the mean temperature at the sea level. Anomalies are referenced to the norm values calculated for the 1996-2016 period. The German Meteorological Institute (DWD[6]) provide forecasts of the anomaly values for two consecutive 3-month periods (2^{nd} to 4^{th} and 3^{rd} to 5^{th} months counting from the current month) for the mean temperatures at 2 m, pressure at sea level and the sum of the precipitation. In this case, norms are calculated for the 1981-2014 period. The American meteorological service (NOAA[7]) generates monthly and seasonal forecasts for a variety of meteorological parameters, including the temperature at 2 m, the sum of the precipitation, geopotential height at 200 hPa and 700 hPa, the temperature at 850 hPa and the sea surface temperature (*SST*) for various regions, including Europe. The system generates forecasts four times a day, based on the initial values gathered from 30 antecedent days. The forecast is generated in three ensembles (40 elements each): E1 includes days 1 to 10 of the 30 days prior to the forecast generation, E2 – days 11 to 20 and E3 – days 21 to 30. Anomalies are calculated based on the norms for the period 1999-2010. The 6-month forecast updated every month, found at the website of the International Institute of Science and Climate at the University of Columbia (IRI[8]), is also worth mentioning. This forecast is more interesting in that it offers predictions of the monthly averages of the temperature and the sum of the precipitation. The forecast is presented as the probability estimation for the mentioned parameters of being above, below or in range of the norms for a given period. In this case, norms are defined as the range from 33% and 66% quantiles of the averages calculated over the last 30-year period (currently until 1981-2010).

However, it is not always feasible to use forecast generated by such global institutes for local purposes due to methodological differences, poor resolution, and the exclusion of local conditions from the model. Therefore, IMGW-PIB constantly works on establishing reliable models for a long-range forecast that will account for local conditions. The model described in this article generates predictions of the general character of temperature, investigating whether

[4] http://www.wmo.int/pages/prog/wcp/wcasp/gpc/gpc.php
[5] https://www.ecmwf.int/en/forecasts/charts/catalogue/seasonal_system5_public_standard_2mtm?facets=Range,Long%20(Months)%3BType,Forecasts&time=2018050100,744,2018060100&stats=tsum
[6] https://www.dwd.de/EN/ourservices/seasonals_forecasts/charts.html
[7] http://www.cpc.ncep.noaa.gov/products/CFSv2/CFSv2seasonal.shtml
[8] https://iri.columbia.edu/our-expertise/climate/forecasts/#Seasonal_Climate_Forecasts

the monthly average is below, above or in the 33th to 66th percentile range for the season defined as the period of three consecutive months. A similar forecast, only global, can be found on the NOAA server. The forecast is generated for 10 selected stations in various regions in Poland. All these stations generate measurement series over a long period of time, without breaks.

Separate regression models are determined in each location by means of the partial least squares (PLS) and principal component regression (PCR) methods (Mevik, Wehrens 2007; Mevik et al. 2016) and sparse partial least squares (SPLS) (Chung, Keles 2010a, b), using a series of predictors for the assumed lengths of dependent periods (from 5 to 30 years). The generated models are assessed based on their ability to reconstruct the dependent variable given the adopted criteria and thus, the variables with the highest score are chosen.

Dataset

In the described forecasting model, the anomaly of the average monthly air temperature for the period of three consecutive months is the dependent variable. Quantile values of the NCEP/NCAR Reanalysis project[9], published by NOAA/OAR/ESRL PSD, Boulder, Colorado, USA (Kalnay et al. 1996), serve as descriptive variables – predictors. Figure 1 shows the location of the grid points and stations for which the forecast was generated.

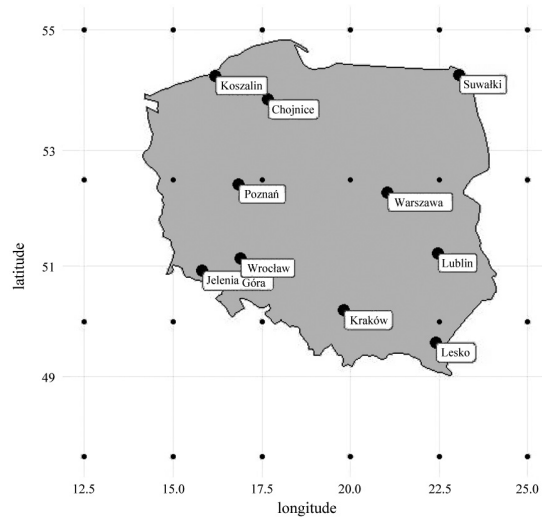

Fig. 1. Location of the stations associated with the generated forecasts; grid nodes indicate points where the values of reanalysis are used as predictors

[9] https://www.esrl.noaa.gov/psd

Predictors

Four variables were chosen as predictors: air temperature (*T*), geopotential height (*HGT*), component of the horizontal wind (*UWIND*) and relative humidity (*RH*). Variables *T*, *HGT* and *UWIND* are generated for each of the 17 levels of geopotential height (from 1000 to 10 hPa), and *RH* is generated for each of the 8 levels (from 1000 to 300 hPa). The NCEP/NCAR reanalysis is published with a 2.5°C resolution. The area of Poland is covered by 24 points, given by longitudes between and including 12.5-25 and latitudes within 55-47.5, as shown in Fig. 1. Each point of the grid covering Poland is represented as 17 values of *T*, *HGT*, *UWIND* parameters and 8 values of the *RH* parameter. Based on the daily data, the 10^{th}, 50^{th} and 90^{th} percentiles are then calculated. This provides information on the dominant and extreme values on each of the parameters in each month, resulting in 4248 potential predictors. It is understandable that not all of the variables are independent, but the regression methods applied here can limit the variable set to only those influencing the predictive variable.

Predicted parameter

The deviation from the long-term average air temperature from 10 selected measuring stations serves as a dependent variable in the model. The stations selected for test calculations are located in Chojnice, Jelenia Góra, Koszalin, Kraków, Lesko, Lublin, Poznań, Suwałki, Warszawa and Wrocław (Fig. 1).

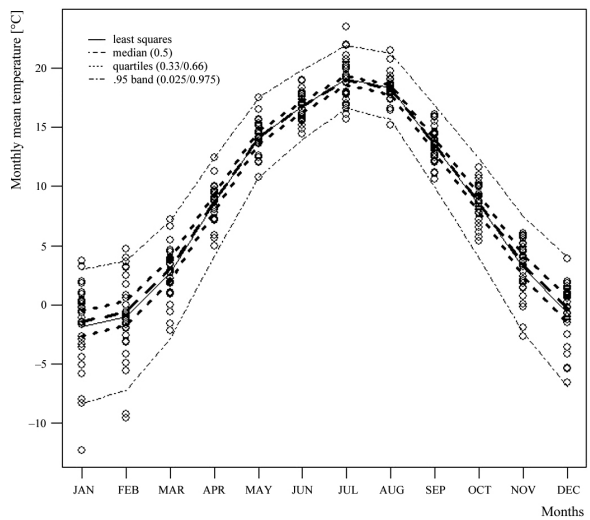

Fig. 2. Characteristics and limits of the norms of the monthly mean temperature for Warszawa

Table 1. The length of the intervals [in °C] of norms (the 33rd and 66th percentiles range) of the monthly mean temperature in selected stations in the period 1981-2010

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Chojnice	2.1	2.1	1.9	1.5	1.2	1.0	0.8	0.9	1.2	1.5	1.8	2.0
Jelenia Góra	2.5	2.4	2.1	1.7	1.2	1.0	0.8	0.8	1.2	1.6	2.1	2.4
Koszalin	1.9	1.9	1.7	1.5	1.2	1.1	0.9	0.9	1.1	1.4	1.6	1.8
Kraków	2.4	2.3	2.0	1.6	1.2	1.1	0.8	0.9	1.2	1.6	2.0	2.3
Lesko	2.5	2.4	2.1	1.6	1.2	1.1	0.9	0.9	1.3	1.6	2.0	2.4
Lublin	2.2	2,1	1.9	1.5	1.2	1.1	0.9	0.9	1.2	1.5	1.9	2.1
Poznań	2.0	1.9	1.8	1.5	1.3	1.1	1.1	1.1	1.3	1.5	1.7	1.9
Suwałki	2.1	2.0	1.9	1.5	1.2	1.1	1.0	1.1	1.3	1.5	1.8	2.0
Warszawa	2.2	2.1	1.9	1.5	1.1	0.9	0.8	0.8	1.2	1.5	1.9	2.1
Wrocław	2.3	2.2	2.0	1.6	1.3	1.1	0.9	1.1	1.3	1.6	2.0	2.2

Here, the average air temperature is defined as the median calculated for the years 1981-2010, and the norm is defined inside the 33rd and 66th percentiles range for the parameter.

Ultimately, the forecast comes down to whether a predictive value will either fit within the normal range or will exceed the defined limits. The calculated ranges are, in general, within 2°C of the norm (Table 1). The summer months have narrower normal ranges than the winter months. For the months of June until September, the range of the norm does not exceed 1.5°C, while for the cold months of January-February and November-December, it often exceeds 2°C. As a result, the forecast error has to be limited to a maximum of 2°C for an accurate forecast of the general character of the average monthly temperature (below the norm, in the norm, above the norm).

This poses very strict requirements on the constructed regression model. Given the number of available potential predictors, the methods used separately take into account the specifics of each location. Additionally, considering a broad range for the time series (from 5 to 30 years up until the first of the forecasted month) accounts for the possible dependence of the current state of the atmosphere on events in the near and distant past. In other words, the length of the considered measurement series of a dependent material varies from 60 to 360 consecutive months.

Method description

Long-range forecasts can be generated using dynamic methods, i.e. by solving the equations describing the atmosphere. Nonetheless, due to the limitations of numerical methods when it comes to the deterministic chaos, the construction

of the statistical model seems to be a better solution. A well-constructed, specialized statistical model, e.g. a multiple regression model, can take into account the impact of many factors, both temporal and physical. However, multiple linear regression has a number of constraints that make it impossible to construct an effective prognostic model.

To properly model the atmospheric events that have an influence on the temperature, the parameters that describe such processes must be included. One of the limitations of the linear regression is that the number of prognostic functions is limited by the minimum of the numbers of dependent and the numbers of independent variables. This places unrealistic restrictions on the number of observations, with most of the literature suggesting as much as 15 to 20 observations per one variable in the model. Furthermore, temperature, geopotential height, and humidity do not entirely satisfy the assumption of independence. This makes it even more difficult to use all the required parameters at the same time in the traditional regression models. Moreover, constructing a regression model by successively adding and removing potential predictors based on their correlation with an explanatory variable does not always guarantee the selection of the optimal model.

Least-squares partial regression methods are free from these constraints. They serve as a convenient extension of multiple linear regression methods, allowing for the inclusion of only a few predictors. Rejecting the independence constraint allows for the inclusion of dependant variables that are highly relevant to the atmospheric state. The described prognostic model uses two methods from the group of iterative partial regression methods: PLS, based on the decomposition of Hermann Wold (Wold 1985; Henseler et al. 2009) and SPLS (Chung, Keles 2010a), which is an extension of the former method. Both are characterized by the occurrence of hidden layers and hidden variables of the model, taken as linear combinations of the predictors. Regression dependencies are determined for these hidden variables. For the SPLS method, the hidden variables are determined on an incomplete set of predictors and a continuous process is used to match the variables to the model. Two R (R Core Team, 2017) packages, the pls and spls packages, were used to generate the model. The pls package (Mevik, Wehrens 2007; Mevik et al. 2016) implements the first method, while the spls package (Chung, Keles 2010b) implements the SPLS method. The regression model was constructed separately for each of the ten stations shown in Figure 1. The calculations were carried out for the different periods from which the coefficients of the regression equation were calculated, from 5 to 30 years immediately preceding the forecasting period. The number of intermediate layers was also variable. In the case of the PLS method, it ranged from 1 layer to the maximum number of layers determined by the training set. For the SPLS method, the number of layers ranged from 15 to 45. For the given parameter set, the optimum model was chosen from the results of the cross-validation,

based on the randomly selected test set. This allowed for the derivation of a set of models for each method. The models were later assessed on the training dataset and those that satisfied the conditions below were then used to generate temperature forecasts for the following three months:

1. Models with an error rate of the dependent variables less than or equal to 10%.
2. Models with more than 70% of all dependent variables within a range of 0.5°C of the true values.

Results

The described prognostic model was tested on the period of three years. The average monthly temperature was predicted for the three consecutive months of February-April 2014 to January-March 2017. Therefore, verifiability analysis of the forecasts was carried out on 360 cases – 36 tests for each of the stations shown in Figure 1. For the PLS method, a set of 78 forecasts (26 dependent periods and 3 sets of procedure parameters), and for the SPLS method a set of 182 forecasts (26 dependent periods and 7 values of the intermediate layers) were generated. The second constraint (of more than 70% dependent variables belonging within the 0.5°C range of true values) was on average met by 10 forecasts generated by the PLS method (from 10% to 19%) and by 104 forecasts generated by the SPLS method (from 52% to 62%). The location and length of the forecast did not seem to influence the fraction of forecasts that met the test conditions.

The selected set of the forecasts was then used to predict the overall character of the mean temperature (*TS*) – estimate its location relative to the norm, i.e. whether it will be within, below or above the normal range. For this, four methods were applied:

- M1 – the mean *TS* value for the set of forecasts was calculated and the obtained value was compared against the normal range for a given month.
- M2 – the median *TS* value for the set of forecasts was calculated and the obtained value was compared against the normal range for a given month.
- M3 – the mean *TS* value was calculated for those values of the forecast set that fit within the 25^{th} and 75^{th} percentile, and the value thus obtained was compared against the standard range for a given month.
- M4 – the value generated from each forecast was classified into within, below or above the normal range class; occurrences of each class were counted and the most frequent was assigned as the output of a given set of forecasts.

The results obtained were also analysed in terms of the dependence of the end month of the training data belonging to a warmer (months 04, 05, 06, 07, 08, 09) or colder (months 01, 02, 03, 10, 11, 12) part of the year. Table 2 presents the results in terms of the warm and cold periods. The percentage of accurate forecasts above 50% for both the whole year and the warm

half-year, and those over 45% for the colder part of the year, were highlighted. Using the M1 method to calculate the predictions resulted in values closest to the true values. This is true for both the whole dataset and for forecasts calculated separately for the warm and cold half-years. The percentage of accurate forecasts generated for the warm half-year was on average higher by 5% than for those generated for the whole year. The regression model gave less accurate forecasts in the cold half-year.

The overall performance of the forecasts seems to be dependent on the location of the measuring station. Most accurate results were achieved for the Warszawa and Wrocław stations, for which the percentages derived from the M1 algorithm for all seasons, ensembles and consecutive months (apart from 2. month in the case of Warszawa) exceeded the threshold of 60%.

Table 2. Results of the forecasts for the whole year and warm and cold half-year periods for the entire analysed data; columns are associated with the scope of the ensemble (separately for each of the regression methods and total) and with the postprocessing method; rows of the table correspond to the consecutive months of the forecast

Method	Ensemble for PLS method				Ensemble for SPLS method				Ensemble for both methods			
	M1 (based on ensemble mean)	**M2** (based on ensemble median)	**M3** (average limited by quantiles)	**M4** (based on the most numerous class)	**M1** (based on ensemble mean)	**M2** (based on ensemble median)	**M3** (average limited by quantiles)	**M4** (based on the most numerous class)	**M1** (based on ensemble mean)	**M2** (based on ensemble median)	**M3** (average limited by quantiles)	**M4** (based on the most numerous class)
Whole year												
1. Month	**51.9**	35.0	35.6	34.4	**51.1**	35.0	34.7	36.7	**51.1**	34.7	34.2	39.7
2. Month	**54.2**	41.7	40.8	41.7	48.9	35.0	35.6	39.7	50.0	35.0	35.6	40.8
3. Month	**54.7**	42.2	42.2	41.4	**55.3**	43.1	43.1	43.9	**55.3**	43.9	43.3	47.8
Warm season												
1. Month	**55.0**	28.9	30.0	30.6	**55.0**	27.2	26.7	30.6	**55.0**	27.2	26.7	32.2
2. Month	**61.1**	44.4	45.0	45.0	**61.1**	44.4	47.8	49.4	**61.1**	45.0	47.2	45.6
3. Month	**61.7**	**56.1**	**56.7**	**53.9**	**60.0**	46.1	46.1	49.4	**60.0**	46.7	46.1	**56.1**
Cold season												
1. Month	**48.9**	41.1	41.1	38.3	**47.2**	42.8	42.8	42.8	**47.2**	42.2	41.7	**47.2**
2. Month	**47.2**	38.9	36.7	38.3	36.7	25.6	23.3	30.0	38.9	25.0	23.9	36.1
3. Month	**47.8**	28.3	27.8	28.9	**50.6**	40.0	40.0	38.3	**50.6**	41.1	40.6	39.4

Conclusion

The presented partial least-squares regression model proves to be a useful tool for determining the type of monthly mean air temperature in the coming three-month period. The predicted value is calculated only based on the model sets selected in the validation procedure. The selected duration of the period preceding the forecast, predictors and the parameters defining hidden layers of models are selected based on the local state (i.e. the correlation between variables). The model allows for the accurate prediction of the type of mean air temperature in relation to the norm for the consecutive 3-month period and can be used in seasonal forecasting. The presented method may be an alternative to using numerical methods in long-term forecasts. Similar models can be constructed for other meteorological characteristics (e.g. precipitation)

Literature

Chung D., Keles S., 2010a, Sparse partial least squares classification for high dimensional data, Statistical Applications in Genetics Molecular Biology, 9 (1), DOI: 10.2202/1544-6115.1492

Chung D., Keles S., 2010b, Sparse partial least squares for simultaneous dimension reduction and variable selection, Journal of the Royal Statistical Society – Series B, 72 (1), 3-25, DOI: 10.1111/j.1467-9868.2009.00723.x

Henseler J., Ringle C.M., Sinkovics R.R., 2009, The use of partial least squares path modeling in international marketing, [in:] New challenges to international marketing. Volume 20: Advances in international marketing, R.R. Sinkovics, P.N. Ghauri (eds.), Emerald Group Publishing Limited, 277-319

Kalnay E., Kanamitsu M., Kistler R., Collins W., Deaven D., Gandin L., Iredell M., Saha S., White G., Wollen J., Zhu Y., Leetmaa A., Reynolds B., Chelliah M., Ebisuzaki W., Higgins W., Janowiak J., Mo K.C., Ropolewski C., Wang J., Jenne R., Joseph D., 1996, The NCEP/NCAR 40-Year Reanalysis Project, Bulletin of the American Meteorological Society, 77 (3), 437-472, DOI: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Mevik B.-H., Wehrens R., 2007, The pls Package: Principal Component and Partial Least Squares Regression in R, Journal of Statistical Software, 18 (2), DOI: 10.18637/jss.v018.i02

Mevik B.-H., Wehrens R., Liland K.H., 2016, pls: partial least squares and principal component regression. R package version 2.6-0, dostępne online: https://CRAN.R-project.org/package=pls (03.09.2018)

R Core Team, 2017, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria

Silingo J., Palmer T., 2011, Uncertainty in weather and climate prediction, Philosophical Transactions of The Royal Society. Series A: Mathematical Physical and Engineering Sciences, 369 (1956), 4751-4767, DOI: 10.1098/rsta.2011.0161

Wold H., 1985, Partial least squares, [in:] Encyclopedia of statistical sciences, vol. 6, S. Kotz, N.L. Johnson (eds.), Wiley, New York, 581-591

S u m m a r y

The article presents the construction of a regression model for the long-range forecast of tercile categories of the monthly mean temperature. Two methods from the group of the partial least squares (PLS) and sparse partial least squares (SPLS) methods were used. The selected methods combine the properties of principal component analysis (PCA) with features of multiple regression methods, and apply the creation of latent layers. These methods also have no restrictions related to the independence of predictors and no constraints on the model dimension. The predictors are percentiles (10%, 50% and 90%) for selected fields of the NCEP/NCAR Reanalysis dataset. The model uses a time series of predictors for periods from 5 to 30 years. The obtained set of forecasts is subjected to the evaluation process based on indicators for the dependent period. This allows for the selection of a reliable ensemble of forecasts. The presented model was tested between January 2014 and December 2016.

Key words: long range forecasts, regression model, partial least squares.