National Research University

Higher School of Economics

Faculty of Informatics, Mathematics, and Computer Science

Department of Information Systems and Technology

PROJECT PROPOSAL

«Analysis of weather forecasting efficiency in our region based on global informational systems of meteorological services»

Student: Golovanov Andrey

Group: 14BI-2

Argument Consultant: Pelinovsky E.N.

Style and Language Consultant: Frolova N.H.

2018

CONTENTS

Abstract

There is a tremendous quantity of numerical methods which are deployed in order to predict the future weather for a specific period. In fact, the weather conditions comprise various characteristics, but it is essential to know air temperature. This is the main metrics of the current research.

The purpose is an independent assessment of weather accuracy for a day, two days and up to six day ahead using three sites: weather.com, gismeteo.ru, meteoinfo.ru. The basic idea is a collection of extreme temperatures during the observational period using the script via Google Apps Script (1st October 2016 – 1st April 2017, 1st October 2017 – 1st April 2018) and calculating the coefficient of determination using real values and previously predicted ones. The final results will be computed via Matlab script.

Moreover, the period was divided differently to obtain more information about the periods which clarify and demonstrate the reason of the particular accuracy. Thus, the most significant deviations of predicted and real extreme temperatures will indicate the dates with some weather accidents.

The average results of the whole period shows 96% of accuracy prediction for one day ahead, 93% for 2 days, 90% for 3 days up to 70% for 6 days. The descending accuracy of scaling days is described quantitatively for various periods.

Key words: weather forecasting, prediction efficiency, meteorological accuracy, forecasting assessment

Introduction

The background of the study

Weather determination is a prerogative of Synoptics as a section of Meteorology which is sufficiently ancient. Historically, the first meteorological stations were erected in 17th century. The next breakthrough was the invention of the synoptic maps in 19th century. Eventually, there were three generally accepted methods of weather forecasting: a synoptic method, a numerical method and a statistical one. Nevertheless, my emphasis is centralized towards the quality of weather prediction.

The issue of weather forecasting accuracy is an extremely vital theme due to the effect of weather relatively to a tremendous amount of businesses. For instance, agriculture has to be aware of the future weather to devise its own strategy which assists to grow as much provision as it is possible. Otherwise, misinterpretation of the weather is able to cost a fortune to the business.

As a rule, the accuracy is assumed to be measured through sophisticated methods and it leads to the probability of accurate values of temperatures, humidity, and wind power on future period. Thus, meteorological services are able to claim these probabilities as quantitative metrics of the weather prediction accuracy.

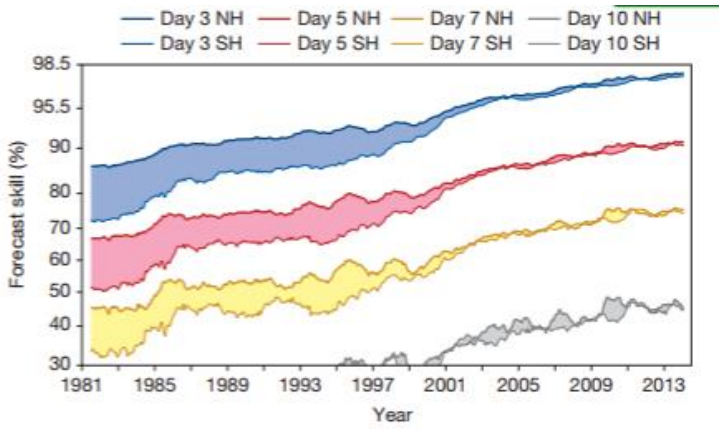

Figure 1

The figure 1 (Bauer Peter, 2015) demonstrates the overall ability of weather forecasting which is considered to be highly accurate in terms of numerical and synoptic methods. Unfortunately, sometimes the Synoptic Meteorology results are able to misinform consumers. Why does it happen?

In spite of it, this accuracy has a lack of self-assessment of present accuracy instead of future one. In fact, only a current stance provides an opportunity to foresee a future event like the weather.

Factually, weather is characterized by various characteristics. Therefore, there has to be chosen one key metric to assess its precision. Temperature is the most essential way to describe the weather as the value significantly affects economy and society.

To recap, the object of my research is the temperature values of weather forecasting meteorological services and the subject comprises the accuracy and the spread of this phenomenon.

Problem statement

The aforesaid inconsistencies lead to the problem of the current accuracy definition. It is determined as:

- The weather accuracy has no self-assessment;
- Self-assessment is based on the previous results to determine a current state;
- There has to be chosen accuracy metrics;
- It is possible to evaluate a limited amount of meteorological services;
- Due to the opportunity to deploy the previous data, it is an obligatory feature for metrics to be statistical ones.

There should be data of different periods, for instance, one day ahead, two days and up to six days ahead. The most appropriate metrics is the coefficient of determination (R-square) and the spread rate as they are based on usage of the previous data.

If it is possible to collect the data from several meteorological databases then R-square demonstrates the current accuracy on various periods. Spread rate assists to emphasize the most extreme values of errors to find the dates and scan them to find the reason of the phenomenon.

The data will be collected in an automatized way and there will be three meteorological services to observe: weather.com, gismeteo.ru, meteoinfo.ru from October to April. The next stage is software processing which has to provide numerical results. It is assumed to be the following:

- Numerical accuracy should decrease if the period will be larger;
- Unstable periods (autumn, spring) should be less accurate;
- The sustainable addition of data will align the fluctuation of values;
- The most extreme errors of specific dates will be relevant to particular event.

All these hypotheses will be proved or refuted via quantitative metrics.

The professional significance

Sometimes we take weather for granted and do not emphasize it as a vital factor of life. Weather conditions are extremely significant to immense amount of businesses. For example, agriculture has to take into account temperature to manage the process of growing different types of cereals. Moreover, thermal stations use the weather prediction results to compute the volume of heat for the proper temperature. Other organizations deploy future precipitation results and etc.

Therefore, my objectives are following:

- To collect data independently from any meteorological services;
- To collect in in automatic way on a regular basis;

- To calculate R-square as the accuracy level;
- To use accuracy level to define what service is more precise;
- To endeavor to clarify, why some dates were predicted too poorly.

If this research is succeeded, it will become a non-ornate tool of choice the best service for any business-user and for meteorologists as well.

Main body

Literature review

It is indispensable for quality of weather prediction to be sufficiently accurate. Unfortunately, current tendency demonstrates that meteorologists endeavor to develop the weather forecasting without an appropriate self-assessment. As a rule, there is no independent organization which is ready to impartially analyze the forecasting precision.

At the current time, weather forecasting is a tremendously developed sphere of a meteorological discipline which is called Synoptics (Zverev Anatoly, 1977). The main purpose of this scientific branch is a permanent exploration of the Earth atmospheric processes which are able to determine the future weather forecast. This forecasting output comprises various information: temperature, humidity, wind direction and force and etc.

The foregoing paragraphs lead to a prevalent purpose of this graduate research. It can be called as an independent assessment of weather prediction accuracy in terms of temperatures. The scope of the research comprises: temperature data assimilation, nonphysical statistics as clue tools, and manually created Matlab scripts.

The previous research had sufficient accent on models for assessment of weather forecasts (Pascal J., 2006) or quality estimation of this accuracy (Mason I., 1982). The assessment methods were substantially objective but prone to be specialized in narrow aspects: wind, magnitude and so on. Due to extremely independent Physics approach, the whole situation was no comprehensive general picture.

That is why, my research is based on Math statistics to generally and independently evaluate the common accuracy of weather prediction deploying automotive temperature data collection and calculation of the mean, variance, coefficient of determination by means of prototyping software. The enlisted set of parameters will be wider and the temperature will contain maximum and minimum values for months from several meteorological databases.

Moreover, modern tendency of articles is devoted to the data assimilation development and sophisticated methods of numerical weather prediction. (Vasilev Pavel, 2015) That is the main gap which has to be overcome: calculation of generally statistic estimators using statistic rows.

Partially, it was made in my previous research, but the current purpose is to extend the set of statistic parameters in order to deploy more complicated ones.

Methodology

The overall quantity of methods can be divided into three groups: program scripts, statistic metrics and numerical methods.

My research imposes an obligation of the data extraction from web sites and its storage in a single file. It is needed in order to process it in a program way. Otherwise, the written script will be extremely redundant in terms of efforts. Moreover, the data from the Internet is immense and homogeneous. This fact would cause a splitting headache if the data were collected manually on a daily basis.

Hereby, usage of Google Application Script (GAS0 as a method of data collection is able to save a tremendous quantity of time resources. In fact, the tool does not imply any payment and the ready script is run on Google's servers without a user's computer expenses. It is beneficial set a time-driven trigger which means how frequent this function should be run. The trigger is able to be customized with determining any period of time. This is the key advantages of the method as it allows parsing data without sophisticated lines of code as often as a user desire, despite the state of a computer.

It seems important to add that the whole data of extreme temperatures will be stored in Google Spreadsheet. It provides several benefits. Firstly, the data is allocated on the cloud, so there is preservation from data loss. Secondly, Google Spreadsheet has an embedded Google Application Script technology that guarantees the full package of compatibility. Thirdly, if there is a force majeure in the case of data deletion then the history of document will assist to recover the data.

The final result will be devised through the written Matlab Script due to the availability of an automatic calculation of needed metrics on the current amount of data almost instantly.

When the data has a consolidated structure then there will be an opportunity to process it via the statistic metrics. Statistics imply the usage of previously accumulated data in order to assess the current situation. This feature is a highly applicable to the problem of the research.

Thus, the evaluation is possible via the coefficient of determination as it demonstrates the level of accuracy which helps to prove or disapprove the hypotheses. Various periods will be calculated analogically. The difference between actual temperature and predicted one will demonstrate the largest spread related to specific dates. It may be a clue for finding the reason of this phenomenon.

To sum up, the input and output will be conducted using prewritten GAS and Matlab Script accordingly. The accuracy and spread will be the supporting indicators for hypotheses evaluation.

Results anticipated

If it is permissible to imitate the output of the initial results then the picture should prove the hypotheses.

Прогноз\Метрика	(R^2)-w	(R^2)-g	(R^2)-m
+1	0,97	0,95	0,96
+2	0,94	0,90	0,93
+3	0,91	0,86	0,91
+4	0,90	0,81	0,84
+5	0,84	0,78	0,80
+6	0,81	0,63	0,70

Table 1

Hereby, data of temperatures was collected with developed Google Application Script in an automatic mode. Therefore, there was no need to collect data manually on a regular basis. The script was written with the usage of JavaScript.

The collected data was stored in Google-tables in order to unbind it from the PC. It was made to save the files from unforeseen factors. Consequently, statistic metrics were calculated through Matlab script, for instance, the coefficient of determination.

This table 1 shows the current results of accuracy using the coefficient of determination. Thus, the accuracy has descending tendency for each of three sites scaling up to six days ahead. The long-term accuracy is sufficiently permissive.

The unsustainable periods demonstrate more approximate precision than winter. The figure 2 and the figure 3 provide information about unaffordable precision on six days ahead for winter period.

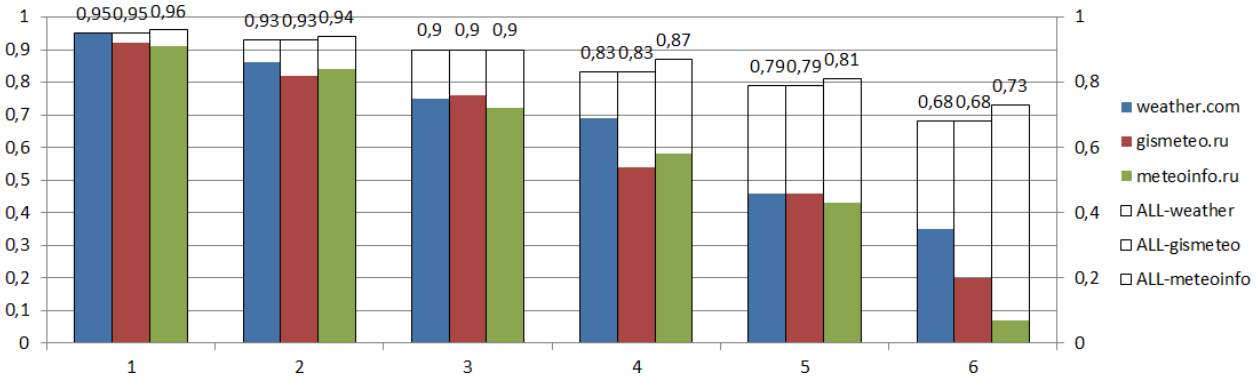

Figure 2

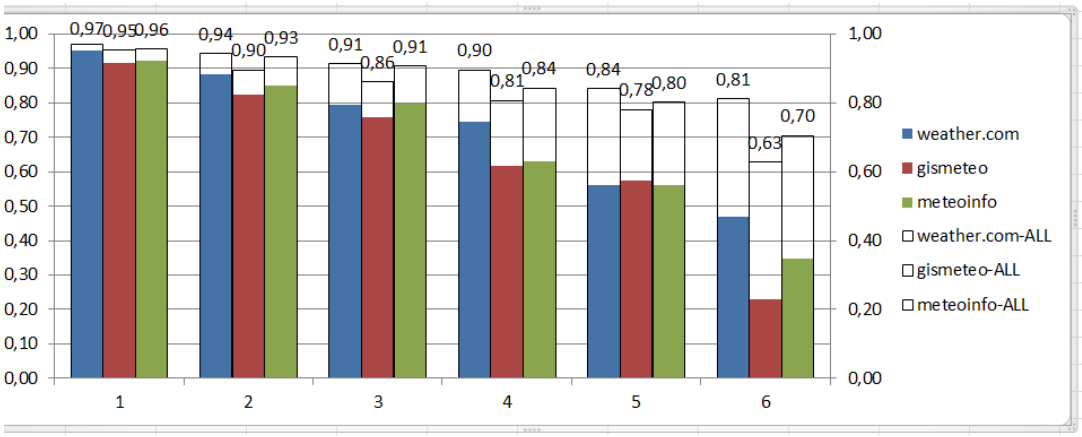

Figure 3

In fact, the figure 2 provides results of observation of three sites from December of 2016th to January 2017th. Consequently, the six-day prediction has only 20% of accuracy. The figure 3 broadens the horizon of observation adding a new period observation which is December of 2017th. The major effect of improvement is represented in 4-6 day prediction than 1-3 day period.

Thereby, the predictive power has more quantitative precision on short-term period which is formed through the coefficient of determination.

Conclusion

To summarize, the devised system of automotive data collection allows evaluating three chosen sites on accuracy for different ranges and periods. There were 1-6 days predictions for autumn, winter and spring periods based on data which was collected during more than a year. The first half of it was transferred manually, another half was obtained in autonomous mode.

The represented system is able to assess the accuracy and the spread of three meteorological stations in terms of maximum and minimum temperatures. It should be added that the comparison of accuracy from three observed sites ambiguously should be correspond to each other. The root is that sites do not show the same temperature in one specific moment as there is different data and places of temperature measurement. That is why, each site has to be assessed independently to figure out its own results.

The whole volume of represented results can be considered as a reasonable assistance for any business user whose goal is defining the most accurate meteorological resource for prediction any of the vital weather characteristics.

Thus, the further research can be transformed to different metrics of weather, for example, wind power or the level of precipitation. It can be done by a meteorological service or anyone else in order to conduct an independent inspection of the weather prediction accuracy.

Reference

1. A. J. Simmons, A. H. (2002). Some aspects of the improvement in skill of numerical weather prediction. *Quarterly journal of the royal meteorological society*, pp. 20-23.
2. Claude Fischer, T. M. (2005, October). An overview of the variational assimilation in the ALADIN/France numerical weather-prediction system. *Quarterly journal of the royal meteorological society*, pp. 20-23.
3. E.N. Pelinovsky, T. T. (2014). *Methods for forecasting natural disasters. Statistic prediction methods Part 1.* Nizhny Novgorod.
4. Mason I. (1982). A model for assessment of weather forecast. *Australian Meteorological Magazine*, pp. 291-302.
5. P. P. Vasilev, E. V. (2015). METHOD FOR PREDICTION PREVAILING MINIMUM AND MAXIMUM OF AIR. 43-57.
6. Q. J. R. Meteorological Society (2005), p. 3.-3. (n.d.). *Meteorological Society*, pp. 3215-3233.
7. Richardson, L. F. (n.d.). Weather Prediction by Numerical Process. pp. 27-30.
8. Stephenson, P. J. (2006). Quality of Weather Forecasts. *Royal Meteorology Society*, pp. 2-5.
9. T. Auligné, A. P. (2007). Adaptive bias correction for satellite data in a numerical weather prediction system. pp. 45-53.
10. Van der Varden (1960). *Mathematic statistics.* STANKIN Press