Model for Processing Historical Timeseries and Establishing Ruleset for Anomaly Detection in Current Sensor Data and General-Purpose Forecasting for Smart Farming in Latvia

Roberts Volkovičs
Modelling of sociotechnical systems
Vidzeme University of Applied Sciences
Valmiera, Latvia
roberts.volkovics@va.lv

***Abstract*—In this article the model is created for establishing ruleset for further anomaly detection in data from smart farming sensors which could be utilized in data cleaning, data analysis and general-purpose forecasting. Model is based on assumption that processes related to smart farming are tied to nature in particular area, for example, Latvia. In case we have historical data for some metric, for example, air temperature, then we can build such a model as long-term knowledge of phenomena provides the basement for such a model to be successful, for example, we know that in Latvia there are four repeating seasons spring, summer, autumn, and winter, but there are even shorter cycles as twelve months with certain known characteristics for each month and day night cycles. Air temperature is more affected by day and night cycles during summer season, less in other seasons.**

Created model will show calculations based on air temperature, but similar approach could be utilized for other metrics which depend on location and seasonality of nature, such as soil moisture level, sunlight, etc. In case, approach is applied elsewhere it is assumed that rulesets are regenerated based on local historical timeseries data.

Keywords — Anomaly Detection, Data Analysis, Forecasting Techniques, Historical Timeseries, Sensor Data, Smart Farming.

I. Introduction

Sometimes researchers perform analysis of timeseries data which come from sensors, for example, sensors of smart farming field. As a smart farming we understand the type of farming that takes an advantage of information technologies in order to maximize efficiency or product yield [1]. As an industry agriculture usually focuses on generation of profit the same way as all the other industries, therefore we can imagine that quite frequently due to the aim of finding the right balance between profit and quality of data and precision of measurements smart farming does not make a focus on quality of sensors, rather it tends to cover fields with many sensors which are affordable. This leads to situation that researchers processing the data need to focus a lot on data cleaning, smoothing and removal of anomalies before the actual analysis of interest is possible.

In this article the model is created which could help smart farming researches to define the ranges of normal data, the ranges of anomalous data and the ranges of most likely data values will take according to long term historical timeseries covering the same seasonality or period so many times that it is possible to utilize statistics and visualization techniques to define the previously mentioned ranges.

Once we have well defined approach and model to process certain case in one location of the world based on historical timeseries of metric, the same could be used elsewhere, based on historical timeseries of that location.

The goal of this research is to build the model allowing to utilize historical time series data to establish ruleset for classifying data points as anomalous or normal, therefore we can state that the hypothesis of this research is that it is possible to build a computational model based on historical timeseries data allowing to classify data points in anomalous and normal, and allowing to predict the most

Online ISSN 2256-070X
https://doi.org/10.17770/etr2025vol2.8569

likely data values for season based on previous values for that season.

II. Materials and methods

As the basement to establish the ruleset to define ranges of normal data and ranges of anomalous data for air temperature in Latvia during particular season, month or day, the open data from 26 Latvian Road Meteo Stations [2] are analysed. It covers the period of more than 10 years which is sufficient as it has enough data points to make the model statistically correct and calculation results and assumptions statistically significant.

All 26 of Latvian Road Meteo Stations have similar characteristics:

1. Data points with values of air temperature are collected each minute;
2. Some data points have no value, some have anomalous value due to sensor error;
3. Overall data quality of 26 Latvian Road Meteo Stations is much higher than quality of data provided by sensors from the fields of smart agriculture;
4. Removing anomalous data and aggregating the rest to closest hour, and by using linear interpolation method from Python Pandas library [3], we can get approximated and correct value of air temperature of approximated hour over territory of Latvia;
5. 20 years of historical timeseries from the same meteo stations could provide statistically correct results for other metrics as well, not only air temperature, but that is out of scope of this research as the focus and aim of the research is to build a model allowing researcher to process such information in standardized way.

As a computational platform the Python programming language will be used focusing on such Python libraries as matplotlib [4], numpy [5], pandas [3], plotly [6], statsmodels [7].

This research will focus on visualization techniques as proof for calculation results as millions of data points will be processed to achieve the aim of the research and there is no other way to demonstrate the proof of this model working other than visualization.

It is assumed that readers can use techniques from statistics and probability behind the scenes on intuitive level as verification of results of visualization.

III. Results and discussion

As the first step, the data were processed to gain air temperatures in Latvia over the previous 3-year period. Calculation result could be seen in Fig. 1 bellow.

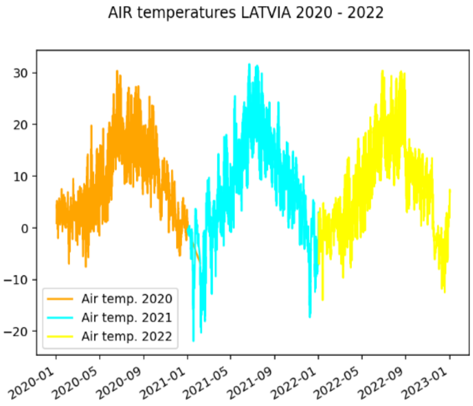

Fig. 1. Air temperatures in Latvia years: 2020, 2021 and 2022.

Fig. 1 demonstrates, that the previous 3-year period of data proves that air temperature data have periodicity over 1-year period and seasonality. Air temperature data do have as well day-night cycles - visualized later during the scope of this article.

Second step, the model allows to adjust the area of most likely values. That could be done by specifying how many timeseries above minimum and bellow maximum values we will consider creating the group of normal data as the rest will create the group of anomalous data, this approach could be compared to Bell-Curve of normal distribution as statistically we can decide the proportion of data point values constructing normal data and proportion of data point values constructing anomalous data.

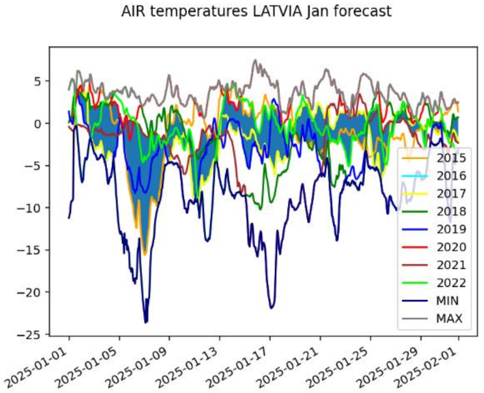

Fig. 2. Air temperature Latvia period from 2015 to 2023.

The Bell-Curve moves over time, like moving average window. That as well means that proportion of data points with normal data is always the same as well of proportion of data points with anomalous data. However, timeseries themselves my cross the border of normal and anomalous data many times during the period of our interest.

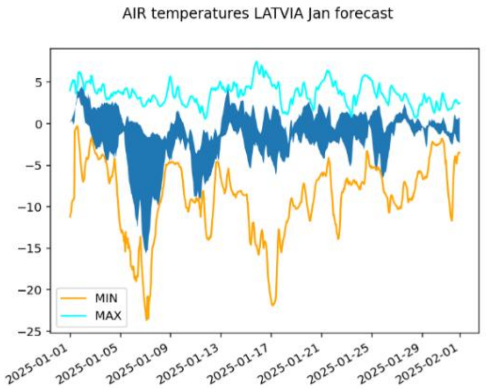

Fig. 3. Air temperature Latvia min/max range most likely range calculated from 2015 to 2023.

As Fig. 2 demonstrates, historical timeseries allow us to create the forecast for January of 2025 utilizing data from period from year 2015 until 2023. For each year data are presented in different colour. The blue area always contains several timeseries inside the area and some timeseries outside the area, as previously mentioned the amount of timeseries inside the area and the amount of timeseries outside the area could be parametrized, but the area itself shows the range of values air temperature most likely should be, based on historic values and proportion of values inside the area is static.

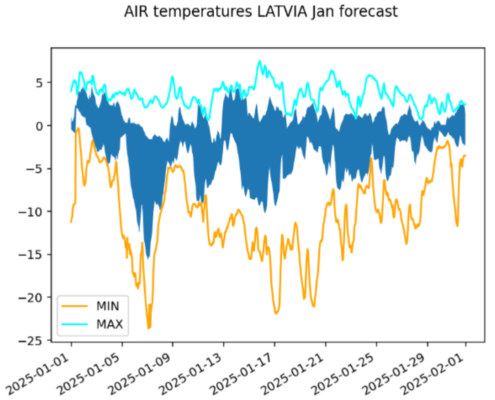

Fig. 4. Air temperature Latvia min/max range most likely range calculated from 2013 to 2015.

Additionally, to values of timeseries from particular year and the area of most likely values, the Fig. 2 shows us the minimum and maximum values based on historical data. Most likely, the values outside the range from min to max could be considered as novelties our outliers, depending on the fact how much historical data we have.

As more historical data we add to the model as more precise the model gets with tendency to expand the range from min to max and expand the range of the most likely values.

In Fig. 3 we see the same data as in Fig. 2, just individual timeseries are removed, but the range from min to max values and the range of the most likely values remains in the graph.

In Fig. 3 the graph is computed based on the period 2015 – 2023 for January, this can be used as a forecast of air temperature based on historic data.

Therefore, we have adjusted the values to timeline to the year of 2025. The same will be done for all the other graphs in this article.

As we see in Fig. 4, after adding a couple of years of timeseries data we slightly expand the range from minimum to maximum and we slightly expand the range of the most likely values.

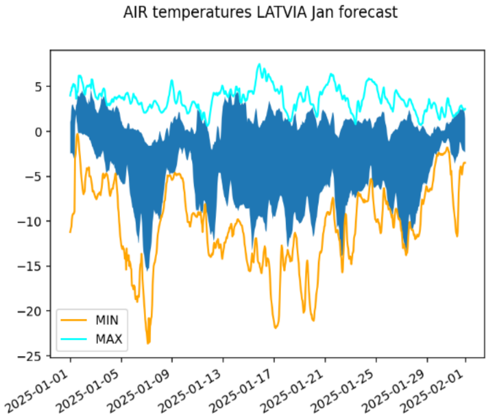

Fig. 5. Air temperature Latvia January min/max range most likely range calculated from 2011 to 2023.

For the scope of this model the decision was made to utilize data from the period of the beginning of the year 2011 to the beginning of the year 2023, for the generation of the ruleset data. In such a way generated graph for January is visible in Fig. 5.

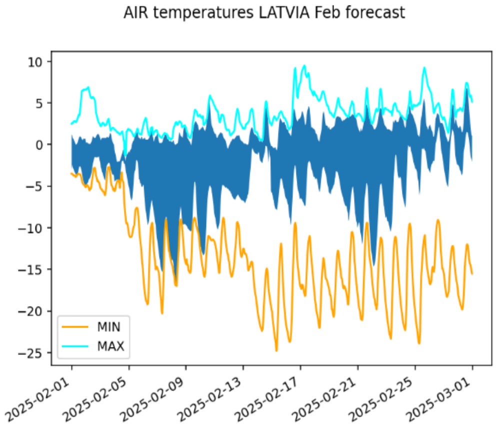

Fig. 6. Air temperature Latvia February min/max range most likely range calculated from 2011 to 2023.

On a modern computer computation necessary to generate the graph takes approximately 5 minutes, therefore it makes sense in the later phase to create some type of aggregation of the results which could be used further by smart agriculture needs.

In such a way complexity of model could be hidden behind the scenes and the work could be reutilized.

In Fig. 6 minimum to maximum range temperatures for February are shown.

To make the results complete, the graphs for the rest of months of the year will be computed.

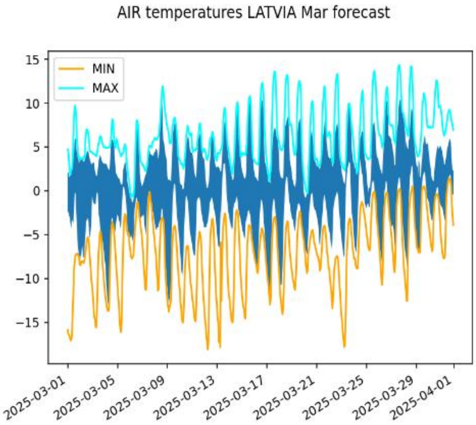

Fig. 7. Air temperature Latvia March min/max range most likely range calculated from 2011 to 2023.

In Fig. 7, the minimum to maximum range temperatures for the month of March can be found.

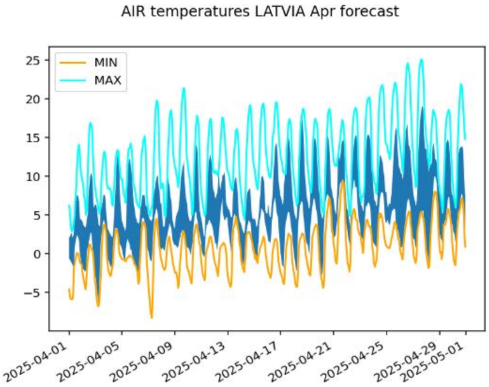

Fig. 8. Air temperature Latvia April min/max range most likely range calculated from 2011 to 2023.

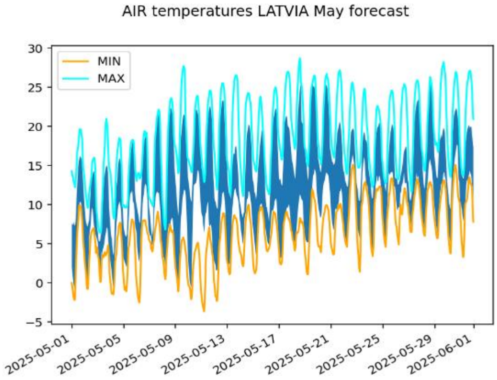

Fig. 9. Air temperature Latvia May min/max range most likely range calculated from 2011 to 2023.

In Fig. 8 minimum to maximum range temperatures for April are visualized.

In Fig. 9, one can find the minimum to maximum range of temperatures for May.

In Fig. 10, the graph of minimum to maximum range temperatures from Jun is shown.

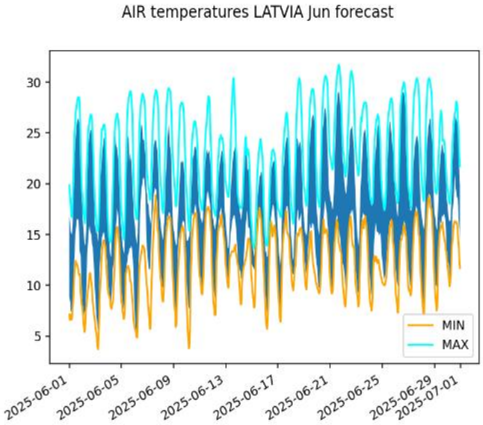

Fig. 10. Air temperature Latvia June min/max range most likely range calculated from 2011 to 2023.

In Fig. 11 minimum to maximum range temperatures for July are visible.

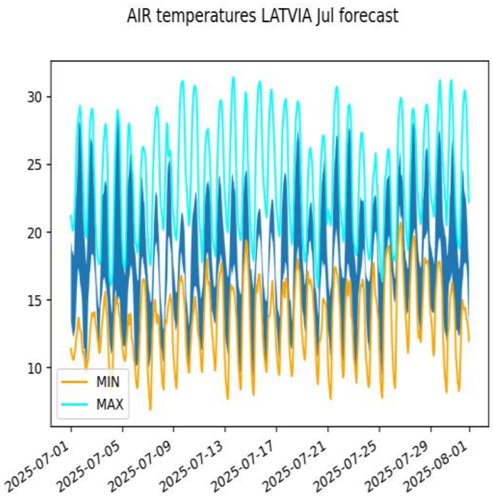

Fig. 11. Air temperature Latvia July min/max range most likely

In Fig. 12 minimum to maximum range temperatures for August can be seen.

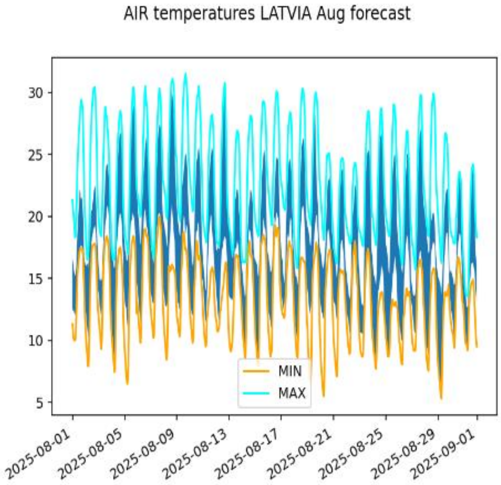

Fig. 12. Air temperature Latvia August min/max range most likely range calculated from 2011 to 2023.

As previously described, air temperatures are very strongly affected by the day/night cycles from March to middle of October. See, Fig. 7 – Fig. 14.

In Fig. 13 minimum to maximum range temperatures for September can be seen.

In Fig. 14 the graph of minimum to maximum range temperatures for October is presented.

In Fig. 15 one can find minimum to maximum range temperatures for November.

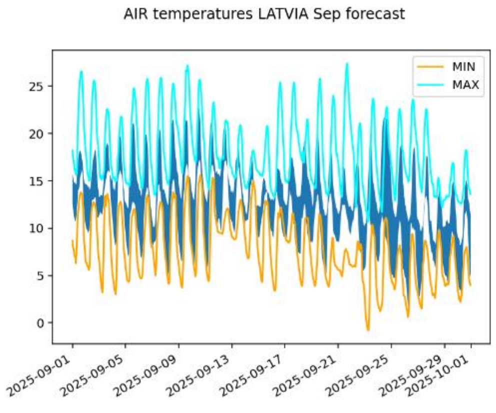

Fig. 13. Air temperature Latvia September min/max range most likely range calculated from 2011 to 2023.

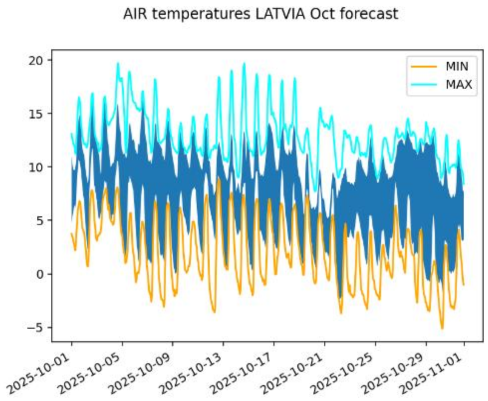

Fig. 14. Air temperature Latvia October min/max range most likely range calculated from 2011 to 2023.

In Fig. 16 minimum to maximum range temperatures for December are visualized.

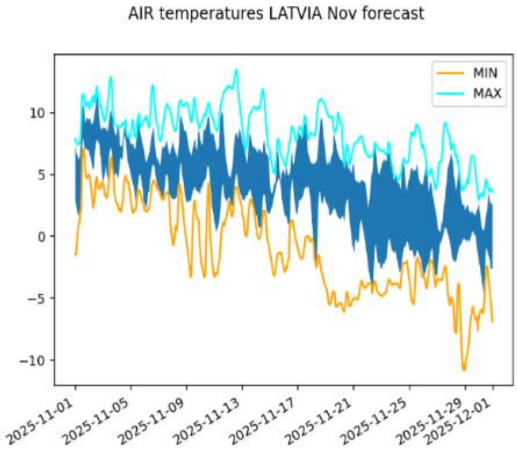

Fig. 15. Air temperature Latvia November min/max range most likely range calculated from 2011 to 2023.

As we see in Tab. 1, it is possible to perform visual analysis or mathematical analysis of long-term historical timeseries and create a parameter table which could be used in further data cleaning, anomaly detection and removal, and forecasting in different real-world smart agriculture scenarios processing output from huge number of low-cost sensors.

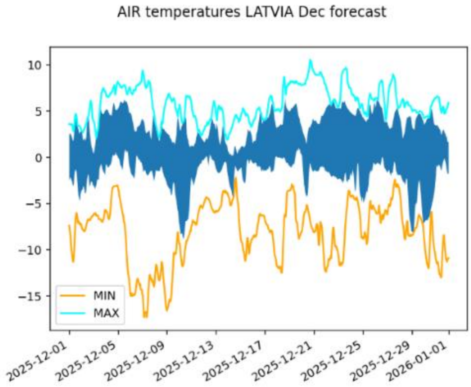

Fig. 16. Air temperature Latvia December min/max range most likely range 2011 – 2023.

TAB. 1 AIR TEMPERATURES MIN/MAX RANGE MOST LIKELY RANGE BY MONTH

Air temperatures MIN, MAX, Expected range by Month				
Month	**MIN**	**MAX**	**Exp. LOW**	**Exp. HIGH**
January	-23	7	-15	4
February	-25	7	-15	5
March	-15	12	-12	10
April	-6	22	-5	15
May	-3	25	0	23
June	4	32	5	25
July	5	32	8	26
August	5	30	7	28
September	0	25	3	22
October	-5	19	-2	15
November	-10	14	-3	10
December	-16	10	-8	6

Main parts of the Python code utilized to generate all the figures from Fig. 1 to Fig. 15 are demonstrated below.

```
...
...
# Calculation of INNER area MIN
# calculated from timesries 2011 - 2022
innerMinJan = at_Jan.groupby('datums').
  apply(lambda grp: grp.nsmallest(3,
'AT').max())
...
...
# Calculation of INNER area MAX
# calculated from timesries 2011 - 2022
innerMaxJan = at_Jan.groupby('datums').
  apply(lambda grp: grp.nlargest(3, 'AT').min())
...
...
# Draw the MIN January line:
# calculated from timesries 2011 - 2022
plt.plot(minJan, c=line_color_Jan, label='MIN')
line_color_Jan = line_color_list_Jan.pop()
...
...
# Draw the MAX January line:
# calculated from timesries 2011 - 2022
plt.plot(maxJan, c=line_color_Jan, label='MAX')
...
...
# Draw the INNER January AREA:
# For example INNER 8 timepoints out of 12
# calculated from timesries 2011 - 2022
plt.fill_between(X, Y11.flatten(),
Y21.flatten())
plt.legend()
...
...
```

Fig. 1. to Fig. 15 were generated using Python library “matplotlib” utilizing interface “pyplot”: “plt.plot” and “plt.fill” calls and calculations were done utilising Python Pandas mathematical functions, grouping, etc.

IV. CONCLUSIONS

During this research we have proved that it is possible to create a model that could be used for low-cost smart agriculture sensor data cleaning and anomaly detection utilizing the method of long term historical timeseries data processing.

It is frequently possible to find existing data in the area of interest which could be treated as more reliable than data from smart agriculture sensors and utilize that to compute parameter table according to this model which later could help with cleaning of data from smart agriculture sensors.

ACKNOWLEDGMENTS

This research was funded by the European Commission, Research Executive Agency grant number 101079206, “Twinning in Environmental Data and Dynamical Systems Modelling for Latvia” (TED4LAT).

REFERENCES

[1] Wikipedia, “Smart farming,” [Online]. Available: https://en.wiktionary.org/wiki/smart_farming [Accessed: Jan. 22, 2025].

[2] Latvian Government Open Data, “Historical data of road weather stations 2001-2020,” [Online]. Available: https://data.gov.lv/dati/dataset/celu-meteo-staciju-vesturiskie-dati-2001-2019-gadi [Accessed: Feb. 09, 2025].

[3] Pandas, “Python Pandas – data analysis library,” [Online]. Available: https://pandas.pydata.org [Accessed: Feb. 09, 2025].

[4] Matplotlib, “Matplotlib: Visualization with Python,” [Online]. Available: https://matplotlib.org [Accessed: Feb. 09, 2025].

[5] NumPy, “The fundamental package for scientific computing with Python,” [Online]. Available: https://numpy.org [Accessed: Feb. 09, 2025].

[6] Plotly, “Plotly Open Source Graphing Library for Python,” [Online]. Available: https://plotly.com/python [Accessed: Feb. 09, 2025].

[7] Statsmodels, “Statistical models, hypothesis tests, and data exploration,” [Online]. Available: https://www.statsmodels.org/stable/index.html [Accessed: Feb. 09, 2025].