Project: Soil moisture Estimation In Estonia Using ISMN Data

Team members: Artur Heimola, Kert Moist, Kaur Kullamäe

Github

TASK 2.

Soil moisture plays a crucial role in agricultural productivity, water resource management and environmental monitoring. It directly affects crop growth, irrigation efficiency and drought assessment. In Estonia, soil moisture measurements are limited in spatial coverage and temporal continuity, which restricts accurate large-scale monitoring. This project addresses this gap by using existing European soil moisture data from the International Soil Moisture Network to build predictive models capable of estimating soil moisture for Estonia based on meteorological variables such as precipitation and air temperature.

This project does not serve a traditional business, the primary beneficiaries are Estonian farmers, environmental researchers and public sector institutions involved in land and water management. The main goals are:

- Develop an accurate and transferable method for estimating soil moisture in Estonia, using indirect data sources.
- Provide insights that support irrigation planning, drought monitoring and sustainable agricultural practices.

Business success criteria

- The models achieve acceptable prediction accuracy, measured using metrics such as MSE, MAE
- A consistent daily soil moisture time series for Estonia is produced

Inventory of resources

- ISMN database providing European soil moisture observations
- Precipitation and air temperature datasets for selected European countries
- Estonian Environment Agency data used to estimate Estonian soil moisture parameters
- Personal computer
- Python
- Machine learning libraries
- Time

Requirements, assumptions and constraints

- Requirements: Reliable (long-term) soil moisture, precipitation, and temperature data from ISMN and Estonian Environment Agency for model training and validation.
- Assumption: Climatic and soil conditions in selected European countries are quite similar to Estonia
- Assumption: Precipitation and air temperature are strong predictors of soil moisture
- Assumption: ISMN observations are accurate and comparable across stations
- Assumption: Soil moisture behavior is sufficiently captured by short-term temporal dependencies (3 to 7 day sum and averages)
- Constraint: Uneven spatial distribution of ISMN stations may affect model generalisation
- Constraint: Missing (NaN) values or irregular measurements
- Constraint: Differences in instrumentation and calibration between stations and/or countries

Risks and contingencies

- Risk: Poor transferability of models to Estonian conditions.
- Risk: Missing low-quality data
- Risk: Overfitting

Terminology

- ISMN – International Soil Moisture Network
- Soil moisture – quantity of water contained in soil
- Random Forest – ensemble model based on multiple decision trees
- LSTM – Long Short-Term Memory model
- MSE - A measure of the quality of an estimator
- MAE - A measure of prediction accuracy

Costs and benefits

Costs:

- Time investment in data collection, data cleaning, processing and modelling
- Computational resource usage

- Data quality assurance

Benefits:

- Improved soil moisture estimation for Estonia
- Support for sustainable agricultural practices
- Once trained, the model can be expanded to new stations, new soil depths or other countries with minimal additional cost

Data-mining Goals

- Train Random Forest / LSTM models using European ISMN data.
- Select and filter countries with enough data
- Evaluate the trained models using RMSE, MAE, R2?
- Apply trained models to Estonian weather data to estimate soil moisture
- The results need to be visualised in a compilable way

Data-mining success criteria

- Models demonstrate stable performance with low prediction error
- Predicted soil moisture values follow realistic temporal patterns.

TASK 3.

Data Preparation and Preprocessing

ISMN data

Data Collection

Training data was obtained from the International Soil Moisture Network (ISMN): https://ismn.earth/en/, covering stations across Europe. Only measurements flagged as "Good" by ISMN quality control were downloaded to ensure reliability.

Data Requirements and Selection

Variables: Precipitation, soil moisture, and air temperature.

Resolution: Hourly measurements, later aggregated to daily averages.

Timeframe: Data initially spanned 01-01-2010 to 05-11-2025. After cleaning, only records from 01-01-2017 onward were retained to reduce volume and ensure consistency.

Coverage: All European ISMN stations.

Completeness: Only stations providing all three variables were included.

Data Availability and Criteria

Stations were screened to confirm complete records for precipitation, soil moisture, and air temperature. Although not all stations have data for the entire calendar year or selected timeframe. Meaning, some stations might have measurement gaps that could not be filled with +- 1 month average.

Included: Stations with all three variables and “Good” quality labels.

Excluded: Stations missing variables or containing lower-quality records.

Data Description

Raw data was provided in .stm format and converted to CSV for processing. Each file contained time series of the variables with hourly resolution, along with metadata such as station ID, location, and timestamps.

Data Exploration and Transformation

Exploration focused on identifying gaps, seasonal variability, and overall distribution.

Hourly data was aggregated into daily averages to reduce noise and resources needed.

Missing values (NaN) were imputed using a ±1 month average fallback, preserving seasonal dynamics.

Using the overall dataset average was avoided to prevent distortion across seasons.

Data Quality Verification

The preprocessing pipeline ensured consistency and reliability.

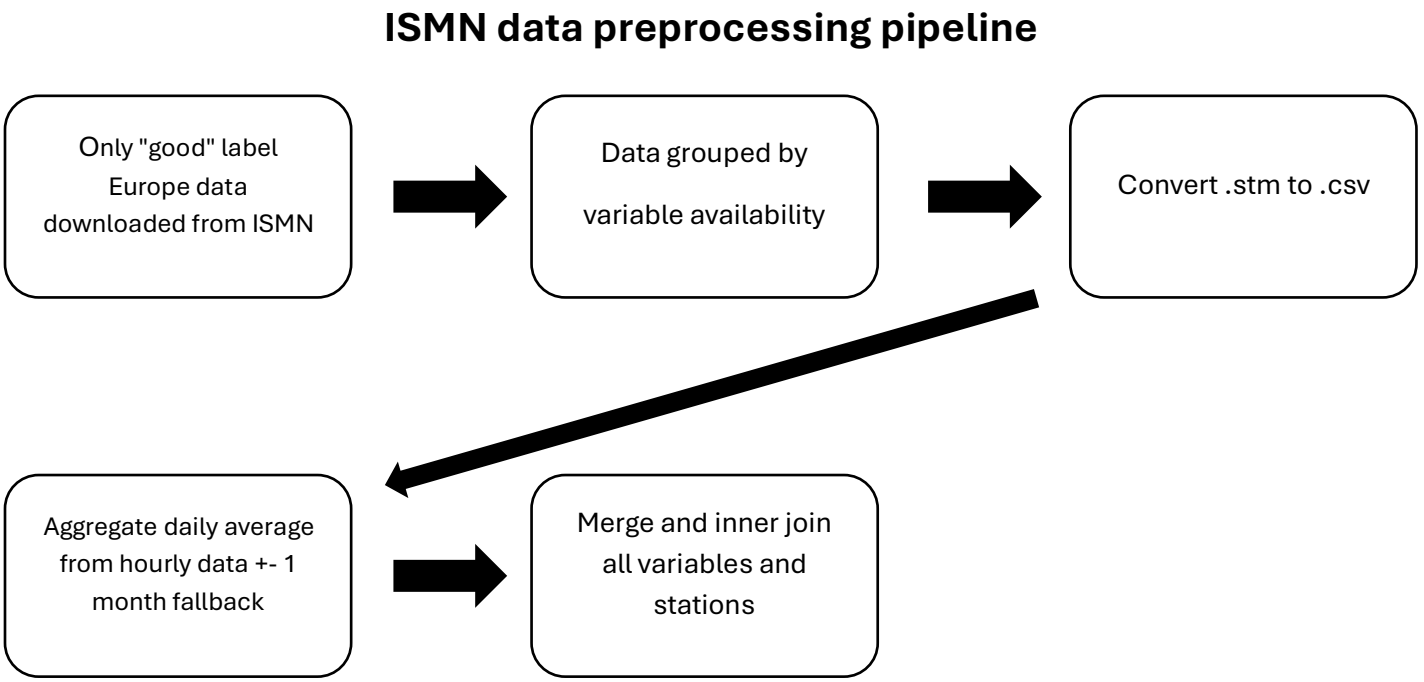
Only “Good” quality measurements were retained.

Seasonal imputation for missing values.

Variables merged from all stations into separate CSV files.

Performed an inner join across stations to align precipitation, soil moisture, and air temperature with corresponding dates.

The final dataset is a cleaned, daily-averaged, multi-variable time series suitable for model training.



Estonian data

Gathering data

The weather data used in this project includes essential meteorological variables such as precipitation and air temperature. As mentioned before, the prediction data for Estonia was gathered from the Estonian Environment Agency's website, under the section of historical weather data.

In Estonia, these weather variables are collected at meteorological stations across the country. For this project, we use measurements only from inland meteorological stations, excluding coastal stations to ensure consistent availability of key variables. The collected

measurements from these stations are available from the mid 2000s up to 2024, with exact starting year varying slightly by station.

Outline data requirements:

- Required temporal coverage is at least 10-15 years for a stable seasonality pattern.
- Data must be available with consistent timestamps across stations.

Verify data availability:

- The required data is open-data and publicly available for all (selected) inland stations.
- Most stations provide continuous measurements from 2004 onward.
- For the end result of this project, the results based on the stations data can be interpolated for other regions across Estonia.

Define selection criteria:

- Meteorology stations that provide continuous precipitation and temperature measurements.
- Long-term historical records.
- Located mostly inland (not coastal) to represent the typical conditions.

Describing data

Data is downloadable in excel (.xlsx) file format and then converted into CSV format. Each station had collected the data in hourly resolution, for all 12 months of the year. The data includes many more parameters which we don't use in this project. **Hourly precipitation observations began at different times depending on the station**.

Exploring data

- Hourly data was aggregated into daily averages.
- For each station, the starting point of the dataset is defined by the time when precipitation measurements first became available.
- Each station's dataset was stored in a separate CSV file.
- All station CSV files follow a unified structure, containing the same variables (day, precipitation (average), air temperature (average).

Data quality verification

- Checking for missing values in precipitation (and air temperature).
- Identifying gaps in the time series for individual stations.
- Ensuring the daily averages were created correctly.
- Validating the consistency of units and formats across the stations.

TASK 4:

Step 1:

Discussion with team, which data to use, how to use, and what tasks each should have.

Initial:

Artur – preprocess ISMN data

Kert – Look into LSTM and Random Forest models

Kaur – Set up GitHub repository

Step: 2

Test models.

Each teammate develops both RF and LSTM models to test out and decide which to use.

In the end, results are compared and model choice is made: Random Forest for simplicity

Step 3:

Combine best features and architecture from each participants RF model (all team members)

- Retrain a single unified RF model using combined settings.
- Perform cross-validation.
- Validate the final model on ISMN stations with data not used in training.

Prepare the model for application in Estonia (all team members)

- Check variable ranges to confirm Estonian climate value fall within European training ranges.
- Run predictions on Estonian stations for full available time range.
- Confirm the plausibility of the results, look for anomalies.

Document, results and presentation (all team members)

- Summarize the results of predicted Estonian soil-moisture data.
- Visualize and compare the predicted Estonian data vs the actual soil moisture data from ISMN.
- Make a map for Estonia with the obtained soil moisture indexes.
- Make a poster upload the .pdf file before the midday of Monday, 8th of December.
- Present the project on Thursday, December 11th, at 14:00-17:00.