Norwegian University
of Life Sciences

Masters Thesis 2024 30 ECTS
Faculty of Science and Technology

A comparative study of soil temperature models, including machine learning models

Mats Hoem Olsen
Master of Science in Data Science

Forword

I would like to thank my advisors and friends. Also the Big Bang for happening.

Glossary

D | H | K | L | M | R | S

D

DataFrame

A table of values. The name is from the python library Pandas used in this study.. 8

H

Hashmap

A list of items where their unique placmnt in the list is detemend by their unique refrence key using a function that maps the key to a placement in the list.. 8

K

Kilden

Norwegian Institute of Bioeconomy Research Kilden. 6

L

LMT

Norwegian Institute of Bioeconomy Research LandbruksMeteorologisk service. 6, 9

Long Short Term-Memory

A Recurent Neural Network with a memory cell to distribute information along the other RNN cells.. 5

LSTM

Long Short Term-Memory. 5

M

MET

The Norwegian Meteorological Institute. 6, 9

MSTL

Multiple Seasonal-Trend decomposition using LOESS. 9

Multiple Seasonal-Trend decomposition using LOESS

Based in the traditional Seasonal-Trend decomposition using LOESS it decomposes a time-series into several seasons, trend, and residual[1].. 9, III

R

Recurent Neural Network

A Neural network that passes information between cells in the same layers.. 5

S

Seasonal-Trend decomposition using LOESS

Takes a timeseries and decomposing it into a trend component, season component, and residual component using local regression for smoothing[2].. 9, III

STL

Seasonal-Trend decomposition using LOESS. 9

Contents

1 Abstract

2 Oppsummering

Keywords: Soil temperature, Machine learning, regression

3 Introduction

In agriculture soil temperature is one of the important parameters to put into consideration when thinking about pest prevention, conservation, and yield prediction. The reasoning for this is that knowing the soil temperature is knowing climate change [3], water management [4], yield [5], nitrogen processes [6] in the soil, calculation of plant-growth [7], when seeds start to sprout [7], potential flooding and erosions[8], and predicting when insect eggs hatch that were laid last winter. Being able to predict the soil temperature into the future will be a huge advantage for farmers, civilians, and scientists.

If it's important, why don't institutions measure it everywhere? There are several reasons for this, but a common reason is that it's expensive to install new equipment on old weather stations. Sometimes the weather station do have the sensors in the fields reading soil temperature at given levels, but due to technical misadventures and unforeseen phenomenons there might be gaps or misreadings that need to be replaced with approximations or NULL values[1]. There are algorithms, models, and statistical tools to interpolate these missing values but they have their drawbacks. For instance approximation by global mean, which is a common method used in timeseries[9]. This method is preserved global statistics, however does not represent local changes. Further more for a good estimation of soil temperature it is useful to include exogenous[2] features.

There has been done research into heat conductivity in soil that has lead to differential equations[10], however these equations[10, 11] are computationally expensive and difficult to simulate, or calculate[6]. To add to the complexity the heat dynamics change depending on soil temperature

In this study 4 methods will be compared and evaluated for the sake of further research into interpolation of missing data in northic countries based on as few features as possible. This study has chosen 2 types of models; Analytical, and Data-Driven models. There will also be base models to compare against, one for each model type.

4 Norwegian introduction

I landbruket er jordtemperatur en av de viktige parametrene å ta i betraktning når man tenker på skadedyrforebygging, bevaring, og avlingsprediksjon. Begrunnelsen for dette er at å kjenne til jordtemperaturen er å kjenne til klimaendringer [3], vannforvaltning [4], utbytte [5], nitrogenprosesser [6], potensielle overfloder of skred[8], plantevekst [7], når frø begynner å spire [7], og forutsi når insektegg klekkes som ble lagt sist vinter. Å kunne forutsi jordtemperaturen inn i fremtiden vil være en stor fordel for bønder, og forskere.

Hvis det er viktig, hvorfor måler ikke institusjoner det overalt? Det er flere årsaker til dette, men en vanlig årsak er at det er dyrt å installere nytt utstyr på gamle værstasjoner. Noen ganger har værstasjonen sensorene i feltene som leser jordtemperatur på gitte nivåer, men på grunn av tekniske feil eller uforutsette fenomener kan det være hull eller feilavlesninger som må erstattes

[1] These values are different from 0 as they represent "no data" and can't be used to do calculations.

[2] Variable that can affect the model, but is not not directly described by the model.

med tilnærminger eller NULL-verdier[3]. Det finnes algoritmer, modeller og statistiske verktøy for å interpolere disse manglende verdiene, men de har sine ulemper. For eksempel tilnærming ved global gjennomsnitt, som er en vanlig metode som brukes i tidsserier[9]. Denne metoden er bevart global statistikk, men representerer ikke lokale endringer. Ytterligere mer for en god estimering av jordtemperatur er det nyttig å inkludere eksogene[4] variabler.

Det har vært gjort forskning på varmeledningsevne i jord som har ført til differensialligninger[10], men disse ligningene[10, 11] er dyre og vanskelige å simulere eller beregne[6]. Videre på grunn av arten av andre partielle derivater ville den numeriske ustabiliteten være for stor for praktiske midler.

I denne studien vil 4 metoder bli sammenlignet og evaluert for videre forskning på interpolering av manglende data i nordlige land basert på så få funksjoner som mulig. Denne studien har valgt 2 typer modeller; Analytiske og datadrevne modeller. Det vil også være basismodeller å sammenligne mot, en for hver modelltype.

[3] Disse verdiene er forskjellige fra 0 siden de representerer ”ingen data” og ikke kan brukes til å gjøre beregninger.

[4] Variabel som kan påvirke modellen, men som ikke er direkte beskrevet av modellen.

5 Theory

This section discusses the theory behind the models used in the

5.1 Linear regression

The regression model will be for the sake of convenience be expressed as the following expression

$$\left(\vec{F}(\mathbf{A})\right)\vec{\beta} = \vec{y} + \vec{\varepsilon}$$

Where $\vec{F}$ is a vector function with following domain $\vec{F} : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times p}$ where $m, n, p \in \mathbb{N}$, $\mathbf{A}$ is the data in matrix form with dimensions $\mathbb{R}^{m \times n}$, $\vec{\beta}$ is the regression terms, $\vec{y}$ is the target (TJM), and $\vec{\varepsilon}$ is the error from modeling.

This basic model to express the linearity of the components to soil temperature. This will function as the base model for regression models.

The $\vec{F}$ is not important, just that your data is shaped by a function.

5.2 Plauborg linear regression model with Fourier terms

Making a linear regression model for soil temperature sensitive to time without introducing more computational heavy operation would be to introduce features that reflect time. In the paper "Simple model for 10 cm soil temperature in different soils with short grass" the author chose to extend the features from air temperature to include also day of year and the air temperature from those days. This means the following F function that Plauborg used would be

$$\vec{F} := [air_t, air_{t-1}, air_{t-2}, air_{t-3}, \sin(\omega t), \cos(\omega t), \sin(2 * \omega t), \cos(2 * \omega t)]^T$$

Where air_t is the air temperature at time t expressed in day of year, ω is the angular frequency to make the argument of sine and cosine expressed in radians. The sine/cosine elements in the F function represent the variations through the day by fitting $\vec{\beta}$ to the yearly variation. To adapt the authors model to an hourly time unit would be to either

1. Extend the F function to include a larger ω coefficient to reflect hourly oscillations in conjunction with daily fluxiations
2. Refit the Fourier terms with a larger ω coefficient to make the oscillations more representative of daily temperature changes.

The larger coefficient could be expressed as $\pi/12$ while the smaller ω for daily values would be rescaled to $\pi/4380$.

The problem with this approsh would be Fouriers Sine-Cosine series approximation which would suggest that Plauborg's method could be subject to overfitting with addition of more terms. On the other hand it gives us a way to compute the coefficients α_i and γ_i for sine and cosine terms respectively, though it would be more numerically stable with a pseudo-inverse computation or a max log likelihood approach. Need to compute condition number of solutions.

5.3 Rankin's finite difference method of simplified heat flow in snow covered soil

A more direct method based on laws of physics develop by Karvonen involves forming a Finite Difference Method (FDM) around point of interest with simplifications to the equations described in *A model for predicting the effect of drainage on soil moisture, soil temperature and crop yield.*

A team of researchers collaborating with the original author found an algorithm by making simplifications to the general differential equations forming a iterative 2-step procedure seen at the procedure 1.

Algorithm 1: Rankin algorithm

```
**Data:** $D, f_d$
**Result:** $T_Z$
$\alpha_t \leftarrow \frac{\partial T/\partial t}{\partial^2 T/\partial z^2}$;
**for** $t \in T$ **do**
    $T_*^{t+1} \leftarrow T_Z^t + \Delta t \times \frac{\alpha_t}{(2Z)^2} \times (T_{air}^t - T_Z^t)$;
    $T_Z^{t+1} \leftarrow T_*^{t+1} * e^{-f_d \times D}$;
**end for**
```

Where $\alpha_t = K_T/C_A$ is the Thermal diffusivity from Fourier's law in thermodynamics, K_T is average soil thermal conductivity, C_A is the apparent heat capacity, and f_d is the damping parameter that has to be empirically derived however for this study it will be estimated from the data through the following estimation

$$f_d \approx \frac{-\ln\left(\frac{T_Z^{t+1}}{T_Z^t + \Delta t \frac{\alpha_t}{(2Z)^2}(T_{air}^t - T_Z^t)}\right)}{2D}$$

The approximation used in the algorithmn 1 assumes that K_T is not dependend on depth . To make the approximation of α_t more accurate the inclusion of rain (θ) to introduce variation can be approximated with

$$\alpha_t \approx \frac{b_1 + b_2\theta + b_3\sqrt{\theta}}{a_1 + a_2\theta}$$

proposed by Kodešová *et al.*[13][5]. To make the computation easier of this Padé-Puiseux[6] approximation hybrid we will realize that α_t is expressed by

$$\frac{b_1 + b_2\theta + b_3\sqrt{\theta}}{a_1 + a_2\theta} \approx \alpha_t \approx \frac{(T_z^{t+1} - T_{air}) * (2z)^2}{(T_{air} - T_z^t) * \Delta t}$$

Thereby only needing a linear regression of two F-functions; $F_1 = [1, \theta, \sqrt{\theta}]^T$ and $F_2 = [1, \theta]^T$ rather than a three step approximation. This algorithm (algorithm 1) will approximate the following integral

$$T = \int_{t_0}^{t_{max}} \frac{K_T}{C_A} \frac{\partial^2 T}{\partial z^2} dt$$

via a Finite Difference Method, although other methods are possible with higher accuracy[7].Must verify for this case! This study will use the FDM used by the author for the purpose

[5] This representation was not proposed by the author however the linear approximations was proposed to approximate K_T and C_A respectfully. Since $\theta \propto m_w$ we can substitute water content with rain in mm since the area is constant and during all messurement the soil type will be the same, however this would need to be resestimated if a station contains a different soil type as the constant has a wide range of values[13].

[6] Padé Approximation is a of the form $\frac{\sum_{i=0}^{\infty} c_i x^i}{\sum_{j=0}^{\infty} c_j x^j}$ and a Puiseux series is a $\sum_{j=N}^{\infty} c_j x^{j/N}$

[7] For example fourth degree Runge-Kutta method[14] which converges quicker than forward-Euler method or FDM.

of making the results in this study comparable with the study presented in the paper "A simple model for predicting soil temperature in snow-covered and seasonally frozen soil."

For inital values this study are utelizing 2 methods under different assumtions:

$$T_z^0 \approx \frac{k\exp(D)}{1+\exp(D)\times(k-1)} \times T_{air}$$

Where k is $K_T * \Delta t/(C_A * (2Z)^2)$, and D is $-f_d * Snow_{Depth}$. This assumes constant air temperature above a constant layer of snow, though unrealistic since air temperature has a tendensy to change during the day due to solar radiation and other climate factors that can cool down or heat up the air. Another problem is the fact that the snow level ramins the same which is also untrue.

5.4 Long Short Term Memory model

When modeling soil temperature it is important to know the previus hours or days to predict the next timestep, for this a natural selection for a data driven model is a recurent network. This type of network makes prediction based on previus timesteps in the data, however the longer timespan the model takse into account the less important are the erlier timesteps in the data. To combat this there was developt an imporoved model called Long-Short Term Memory model[15] that deploys a memory cell that feeds information from erlier timesteps to the late ones. To make sure that redundant information or unimportant informantion dont get feed forward there are also forgetting gates that removes some of the newly learned patterns and integrates it into the memory cell.

5.5 Attention aware LSTM model

6 Method

6.1 Source of data

For this comparative study the following data sources will be used
1. LMT 2. Xgeo 3. Kilden 4. MET

6.2 Dataset

The dataset is chosen from four regions in Norway; Innlandet, Vestfold, Trøndelag, and Østfold. From each region are four stations picked:

Region	Name	ID	Drain type	Soile category	Texture	MET name	Latitude	Longdetude
Innlandet	Apelsvoll	11	Selvdrenert	CM	17	SN11500	60,70024	10,86952
Innlandet	Fåvang	17	Selvdrenert	CM	15	SN13150	61,45822	10,1872
Innlandet	Ilseng	26	Selvdrenert	PH	17	SN12180	60,80264	11,20298
Innlandet	Kise	27	Vannmettet	GL	99	SN12550	60,77324	10,80569
Trøndelag	Kvithamar	57	Vannmettet	ST	16	SN69150	63,48795	10,87994
Trøndelag	Frosta	15	Selvdrenert	LP	13	SN69655	63,56502	10,69298
Trøndelag	Mære	34	Selvdrenert	RG	14	SN71320	63,94244	11,42527
Trøndelag	Rissa	39	Vannmettet	PL	13	SN71320	63,58569	9,97007
Vestfold	Lier	30	Vannmettet	ST	16	SN19940	59,79084	10,25962
Vestfold	Sande	42	Vannmettet	ST	16	SN26990	59,6162	10,22339
Vestfold	Tjølling	50	Selvdrenert	AR	13	SN27780	59,04641	10,12513
Vestfold	Ramnes	38	Vannmettet	ST	16	SN27315	59,38081	10,2397
Østfold	Rakkestad	37	Vannmettet	ST	18	SN3290	59,38824	11,39042
Østfold	Rygge	41	Selvdrenert	AR	13	SN17380	59,39805	10,75427
Østfold	Tomb	52	Vannmettet	ST	16	SN17050	59,31893	10,81449
Østfold	Øsaker	118	Vannmettet	ST	18	SN3370	59,31936	11,04221

Table 1: Station information from stations used in this study. The texture class is defined in this article: https://nibio.no/tema/jord/jordkartlegging/jordsmonnkart/dominerende-tekstur-i-overflatesjikt

All stations are sampled from the date[8] 03-01 to 10-31 from 2016 to 2020. The features rain (RR), mean soil temperature at 10cm (TJM10), mean soil temperature at 20cm (TJM20), and air temperature at 2m (TM) are sampled from the LMT database. The snow parameter is sampled from MET via Xgeo for imputed values in areas where there are no messured values. The soil type, and soil texture is sampled from Kilden from Norwegian Institute of Bioeconomy Research.

6.2.1 Selection process

The selection process for finding these station can be compiled into these steps

1. Recommendation from Norwegian Institute of Bioeconomy Research
2. Compute the missing values in the data
3. Missing values analyse

[8] Format month-day

FROST	SQL approximate Code
Stations with rain	**SELECT** StationName **FROM** FROST ↪ **WHERE LIMIT** 4
Station ID	**SELECT** StationID, LMTID **FROM** FROST, ↪ LMT **WHERE**
LMT	Code
Meteorological data	**SELECT** ID,**date**,TM,RR,TJM10,TJM20 **FROM** ↪ LMT **WHERE date IN BETWEEN year** ↪ −03−01 **year**−10−31 **AND** ID = LMTID

Table 2: What was requested from the varius databases that was used in this study.

4. Searching LMT database for alternative station candidates if current data is insufficient
5. If some station was replaced the repeat step 2

The plots of stations follow a simple representation where the y-axis represent the year and the x-axis represent the index of the data as all tables are taken from the same period. A circle represent a singluar na values, while a band represent a series of 2 or more missing values. The colours represents the features used in this comperative study. This representation of the missing values will indicate sesonal, and systematic removal of data and give an overall indication of how much data is missing. To get further insight into the data a report is generated in parallel to the plots describing precise date and time of all values and which other parameter values is also missing values in the same period. See appendix **??** for the full detail of the report generation and appendix A for na-plots of the station chosen for this study.

6.2.2 Collection of data

The method used was a powershell[9] script that called the respective institutions servers using the "curl" program[10] to send an http request for the timeseries starting from 2014 to 2020 in the interval 1 of May to 31 of October. Code for data collection can be viewed in appendix **??**. The data is stores as an either a csv file or a json file for easy retrieval and manual control of values.

6.2.3 Labeling of stations between Nibio and MET

Since Nibio and MET have different names for the same stations one must compile a list that converts Nibio ID to MET ID. This was performed with these requests Where ID is the Nibio Id for the given station, Frost.ID is the MET id, ID.latitude is the latitude gathered from Nibio, ID.longitude is the longitude gathered from Nibio. These variables can be swaped out for the relevant station.

6.2.4 Storage of data

The storage of the data is done through two data structures; Hashmap and DataFrame from the package pandas. The transformation of data is done with a costume datatype called "DataFileHandler" which is converted to a module for convenience. The keys for the hashmap is chosen by the naming of the data files and the pattern given to the class. To escalete modeling the data will also be exported to a binary file for faster retrieval.

[9] Version 7.3.11

[10] curl 8.4.0 (Windows) libcurl/8.4.0 Schannel WinIDN

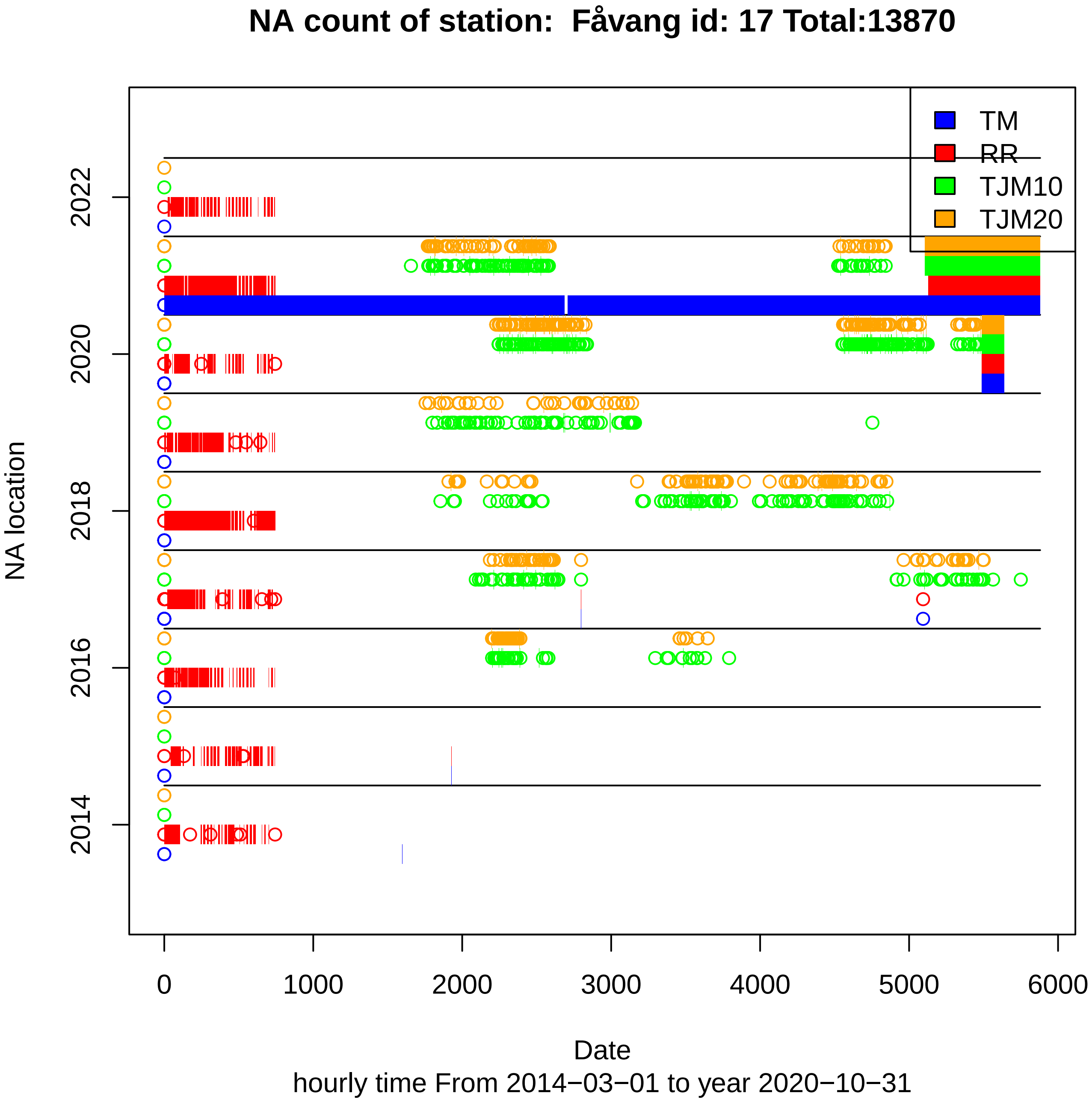

Figure 1: Visual representation of missing values at station 17 from 2014 to 2022

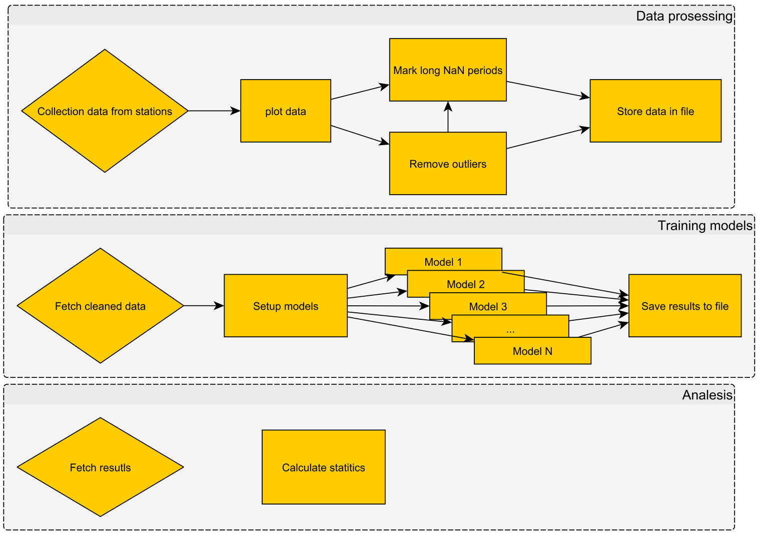

Figure 2

The data structure used to store the data from the different stations is called "DataFileHandler" and stores the data in a tree-structure where indexes are dictated by the filename. It has several built-in functions to assist with data partitioning, and merging of data. This makes it easier to move and store all 846 720 observations from 16 station from 4 regions[11].

6.3 Data cleaning and treatment

To use the data in this study it must be cleaned and treated for training. The following methods were picked common practice in litterateur with new methods based on the decomposition of the data in the from of Seasonal-Trend decomposition using LOESS (STL)[2][12].

6.3.1 Outlier detection and removal

Though the data fetched from LMT is treated and controlled the external data from MET might not be, and this research project incorporated raw, untreated data from LMT to fill inn missing values.

The method to quickly find obvious outliers was to look at the following condition

$$|Z[|\Delta T|]| > \sigma_5$$

This condition looks at the absolute difference between consecutive measurements and calculates the z-score for each observation. It is expected that the change in temperature can't be too rapid.

[11] there are 4 stations per region.

[12] In this study we expand this for multiple seasons using Multiple Seasonal-Trend decomposition using LOESS (MSTL)[1], but the theory of this imputation method remains the same.

6.3.2 Missing value imputation

The data has missing values, in particular during early Fall when there were sub-zero temperatures meaning any rain measurements done during this period would have unpredictable fluctuations since at negative temperatures water can freeze, get clogged up with residual bio-material from the surrounding area

1. linear imputation
2. backwards and forwards first available observation
3. global mean replacement
4. STL decomposition with above methods to impute components

The last method, using STL, was chosen because it would in principle be simpler to impute a less noisy signal than a noisy one. The methods will be tried on 40% of the data, chosen at random to avoid favourable sets that might make the results inaccurate. The reson for only using a fraction of the data rather than all 1.1 millon observations is due to time constrains. For an overview of the prosedure look at the algorithmn 2.

Algorithm 2: Interpolation search

```
Data: Dataset, Methods:(method, configuration)
Result: (method, configureation, length)
for $M \in Methods$ do
    $Results_M \leftarrow \{M, Score : +\infty, Length : 0\}$;
end for
for $M \in Methods$ do
    $R_{method} \leftarrow 0$;
    for $data \in Dataset$ do
        Apply $M$ to all non-missing ranges in $data$;
        Calculate score;
        Add score to $R_{method}$
    end for
    if $\mu(R_{method}) \leq Result_M[Score] \& Length > Result_M[Length]$ then
        Update $Result_M$ with new values;
    end if
end for
Choose smallest Score from $Results$;
```

6.4 Setup of models

The models are set up in according to the relevant paper the model is fetched from, alternatively reuse the code made by the author. When importing the data to the model there will be modifying to the original code to facilitate for the model as far as it goes. Any modifications will be in the appendix under section ??. For the convenience of the reader all code is using the sklearn estimator class to make all the models discuses in this study more user friendly and

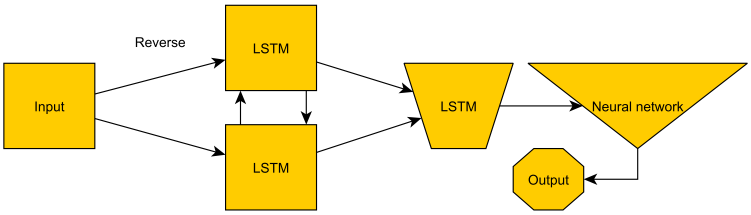

Figure 3

compatible with sklearns other functions. The details of the models will be discussed in section 5, this section discusses the setup and implementation of the models.[13]

6.4.1 Basic Linear model

The linear model (sec 5.1) utilises in the study is created from the python model sklearn (or scikit-learn according to pythons package manager)

6.4.2 Plauborg

The Plauborg regression will be formulated as a linear regression problem so that the LinearRegression function in the Sci-kit module can be used. For the parameters used in the paper[12] the F function defined in section 5.2 will be formulated with loops to give rise 3 more parameters for fine-tuning the model.

6.4.3 BiLSTM

Soil temperatures are dependent on earlier timestemps meaning that to make a good prediction one needs to include temperature from $t, t-1, \ldots, t-k$ to make a decent prediction. As a base model it is usefull to evalueate the data both forwards and backwards to find features that is only notisable in a sertan direction. The BiLSTM defined is crafted with Tensorflow's Keras module for ease of use.

6.4.4 Attention-aware ILSTM

It is a know fact, since 1849[14], that to know previous weather patterns will greatly improve prediction accuracy. To improve the accurasy even more a model can focus one specific patterns that has a big impact on the prediction. The ILSTM attempts to do this by including a new thecnick in data science called attention[16] that takes a collection of data and gives each element a weight associated with importance. When that paper was published it was focused on translation between English and German, however the paper published by Li *et al.* uses this novel teknick to do both time and feature importance and from that make a prediction.

6.5 Metrics

The metrics used in this study are

[13] Caution to the reader; The code used was run on the Linux subsystem (Debian) on windows due to the fact that the current version of tensorflow can't run on Windows.

[14] First weather prediction made by Joseph Henry in 1849

• Mean Squere Error • Mean Absolute Error • Explained Variance • bias • Log Condition number • digit sensitivity

Soil temperatur as a differnet behavor than air temperature since energy (temperature) though the soil gets dampen and delayed. Since the data used in this study has outliers that was not cought during datatreatment, which has been addresed, the author of this study desided to include two more metrics that are not usually included in the evaluation; The log condition number, and digit sensitivity. Both metrics are based on the calculation of the condition number defined as

$$\kappa = \lim_{\varepsilon \to 0^+} \sup_{|\partial x| \leq \varepsilon} \frac{|f(x+\partial x) - f(x)|}{|f(x)|} * \frac{|x|}{|\partial x|} \tag{1}$$

This is not feasible to calculate since infinite calculations with infinitesimal numbers is not possible as per April 24, 2024for simulation approach[15]. Therefore this paper uses algorithm 3 to approximate κ for all the models.

Algorithm 3: Method for calculating κ. $\mathcal{U}$ is a uniform random distrebution in a range.

```
Data: Data
Result: log($\kappa$)
Let $\kappa_f$ be the function 1;
$\kappa \leftarrow 0$;
for $i \in 1 \ldots |Data|$ do
    $\partial x \leftarrow \mathcal{U}_{[-\sqrt{\varepsilon/|Data|}, \sqrt{\varepsilon/|Data|]}}$;
    $k \leftarrow$ calculate with $\kappa_f$ from $x$ and $x + \partial x$;
    if $k > \kappa$ then
        $\kappa \leftarrow k$;
    end if
end for
return $\kappa$
```

The digit sensitivity is included to give an intitive understanding of κ and is computed simply as $\log_e(\kappa)+1$. This number tells us the significant digit generated from the model. If the number is less than 0 then its the ith digit after the decimal point.

For the rest of the metrics, they are defined as follows

- RMSE $= \sqrt{\frac{\sum(y_{\text{pred}} - y_{\text{truth}})^2}{n}}$
- MAE $= \frac{\sum|y_{\text{pred}} - y_{\text{truth}}|}{n}$
- bias $= \frac{\sum(y_{\text{pred}} - y_{\text{truth}})}{n}$
- Explained variance $= 1 - \frac{\sum(y_{\text{pred}} - y_{\text{truth}})^2}{\sum(y_{\text{pred}} - \vec{y})^2}$

Where $\vec{y}$ is the mean of the target, y_{pred} is the predicted data, and y_{truth} is the observed soil temperature.

[15] This calculation is possible for some models, for instance linear regression models when converted to the form $A\vec{\beta} = \vec{y}$.

6.6 Use of Artificial Intelligence in this paper

In this paper there has been used Artificial Intelligence (AI), specifically Bing Chat / Copilot hosted by Microsoft Cooperation with special agreement with The Norwegian University of Life Science (NMBU), for the following purposes:

1. Formalising sentences and rephrasing sentences.
2. Spellchecking
3. Code generation of basic consepts and structures (tree traversal, template for generic classes)

It is important to emphasize that our engagement with AI have been actively curated and verified with known information. All code underwent rigorous manual inspection within a dedicated testing environment. Furthermore, no confidential or sensitive information was shared with the AI; our interactions focused solely on broad topics and general inquiries. To validate the accuracy of AI-generated responses, we cross-referenced them with established research papers and textbooks.

7 Results

poop

8 Discussion

8.1 Future work

The models chosen in this study is not a representative sample of current knowledge of soil temperature modelling, and this study did not aim for optimizing the models beyond what the original authors have already done with the exception for base models used for comparison puposus. A more comprehensive is needed of more complex models that utelises cutting edge technologies, techniques, and theory. One of which is logic based models, for instance ASPER[17] that tries to incoerate logical descriptions of the problem and limits the model for better or equal results based on fewer samples[18]. Another approch is to incorporate randomness into the deterministic models to explain the variation in the data, for instance fractional Brownian motion[19].

9 Conclution

Everything is okay

10 Bibliography

References

[1] Kasun Bandara, Rob J. Hyndman, and Christoph Bergmeir, *MSTL: A seasonal-trend decomposition algorithm for time series with multiple seasonal patterns*, Jul. 28, 2021. arXiv: 2107.13462[stat]. [Online]. Available: http://arxiv.org/abs/2107.13462 (visited on 03/19/2024).

[2] Robert B. Cleveland, William S. Cleveland, and Irma Terpenning, "STL: A seasonal-trend decomposition procedure based on loess," *Journal of Official Statistics*, volume 6, number 1, page 3, Mar. 1990, Num Pages: 3 Place: Stockholm, Sweden Publisher: Statistics Sweden (SCB), ISSN: 0282423X. [Online]. Available: https://www.proquest.com/docview/1266805989/abstract/4DFACD0236364745PQ/1 (visited on 10/11/2023).

[3] Qingliang Li, Yuheng Zhu, Wei Shangguan, Xuezhi Wang, Lu Li, and Fanhua Yu, "An attention-aware LSTM model for soil moisture and soil temperature prediction," *Geoderma*, volume 409, page 115 651, Mar. 2022, ISSN: 00167061. DOI: 10.1016/j.geoderma.2021.115651. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S001670612100731X (visited on 10/05/2023).

[4] Meysam Alizamir, Ozgur Kisi, Ali Najah Ahmed, Cihan Mert, Chow Ming Fai, Sungwon Kim, Nam Won Kim, and Ahmed El-Shafie, "Advanced machine learning model for better prediction accuracy of soil temperature at different depths," *PLOS ONE*, volume 15, number 4, Lei Lin, Ed., page 25, Apr. 14, 2020, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0231055. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0231055 (visited on 09/29/2023).

[5] Ha Seon Sim, Dong Sub Kim, Min Gyu Ahn, Su Ran Ahn, and Sung Kyeom Kim, "Prediction of strawberry growth and fruit yield based on environmental and growth data in a greenhouse for soil cultivation with applied autonomous facilities," *Korean Journal of Horticultural Science and Technology*, volume 38, number 6, pages 840–849, Dec. 31, 2020, ISSN: 1226-8763, 2465-8588. DOI: 10.7235/HORT.20200076. [Online]. Available: https://www.hst-j.org/articles/doi/10.7235/HORT.20200076 (visited on 10/05/2023).

[6] Katri Rankinen, Tuomo Karvonen, and D. Butterfield, "A simple model for predicting soil temperature in snow-covered and seasonally frozen soil: Model description and testing," *Hydrology and Earth System Sciences*, volume 8, number 4, pages 706–716, Aug. 31, 2004, ISSN: 1607-7938. DOI: 10.5194/hess-8-706-2004. [Online]. Available: https://hess.copernicus.org/articles/8/706/2004/ (visited on 03/17/2023).

[7] Cong Li, Yaonan Zhang, and Xupeng Ren, "Modeling hourly soil temperature using deep BiLSTM neural network," *Algorithms*, volume 13, number 7, page 173, Jul. 17, 2020, ISSN: 1999-4893. DOI: 10.3390/a13070173. [Online]. Available: https://www.mdpi.com/1999-4893/13/7/173 (visited on 03/17/2023).

[8] Joris C. Stuurop, Sjoerd E.A.T.M. Van Der Zee, and Helen Kristine French, "The influence of soil texture and environmental conditions on frozen soil infiltration: A numerical investigation," *Cold Regions Science and Technology*, volume 194, page 103 456, Feb. 2022, ISSN: 0165232X. DOI: 10.1016/j.coldregions.2021.103456. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0165232X21002378 (visited on 01/31/2024).

[9] Mathieu Lepot, Jean-Baptiste Aubin, and François Clemens, "Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment," *Water*, volume 9, number 10, page 796, Oct. 17, 2017, ISSN: 2073-4441. DOI: 10.3390/w9100796. [Online]. Available: http://www.mdpi.com/2073-4441/9/10/796 (visited on 02/17/2024).

[10] Tuomo Karvonen, *A model for predicting the effect of drainage on soil moisture, soil temperature and crop yield*. Otaniemi, Finland: Helsinki University of Technology, Laboratory of Hydrology and Water Resources Engineering, 1988, xvi, 215, Open Library ID: OL15197205M.

[11] Jean Baptiste Joseph Fourier and Alexander Freeman, *The analytical theory of heat*. New York: Cambridge University Press, 2009, OCLC: 880311398, ISBN: 978-1-108-00178-6.

[12] Finn Plauborg, "Simple model for 10 cm soil temperature in different soils with short grass," *European Journal of Agronomy*, volume 17, number 3, pages 173–179, Oct. 2002, ISSN: 11610301. DOI: 10.1016/S1161-0301(02)00006-0. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1161030102000060 (visited on 03/17/2023).

[13] Radka Kodešová, Miroslava Vlasáková, Miroslav Fér, Daniela Teplá, Ondřej Jakšík, Pavel Neuberger, and Radomír Adamovský, "Thermal properties of representative soils of the czech republic," *Soil and Water Research*, volume 8, number 4, pages 141–150, Dec. 31, 2013, ISSN: 18015395, 18059384. DOI: 10.17221/33/2013-SWR. [Online]. Available: http://swr.agriculturejournals.cz/doi/10.17221/33/2013-SWR.html (visited on 02/29/2024).

[14] Carl Runge, "Ueber die numerische Aufl sung von Differentialgleichungen," *Mathematische Annalen*, volume 46, number 2, pages 167–178, Jun. 1895, ISSN: 0025-5831, 1432-1807. DOI: 10.1007/BF01446807. [Online]. Available: http://link.springer.com/10.1007/BF01446807 (visited on 02/29/2024).

[15] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, volume 9, number 8, pages 1735–1780, Nov. 1, 1997, ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: https://direct.mit.edu/neco/article/9/8/1735-1780/6109 (visited on 10/18/2023).

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, 2017. arXiv: 1706.03762[cs]. [Online]. Available: http://arxiv.org/abs/1706.03762 (visited on 04/16/2024).

[17] Trung Hoang Le, Huiping Cao, and Tran Cao Son, *ASPER: Answer set programming enhanced neural network models for joint entity-relation extraction*, version: 1, May 24, 2023. arXiv: 2305.15374[cs]. [Online]. Available: http://arxiv.org/abs/2305.15374 (visited on 03/12/2024).

[18] Fadi Al Machot, *Bridging logic and learning: A neural-symbolic approach for enhanced reasoning in neural models (ASPER)*, Dec. 18, 2023. arXiv: 2312.11651[cs]. [Online]. Available: http://arxiv.org/abs/2312.11651 (visited on 03/12/2024).

[19] A. Di Crescenzo, B. Martinucci, and V. Mustaro, *A model based on the fractional brownian motion for the temperature fluctuation in the campi flegrei caldera*, Jul. 20, 2022. arXiv: 2110.13546[math, stat]. [Online]. Available: http://arxiv.org/abs/2110.13546 (visited on 03/12/2024).

A Plots

B Tables

Table

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway