Keene Particulate Matter Project - Roadmap Document

Index

1. [Data Science Lifecycle Review](#_Data_Science_Lifecycle)
2. [Problem Definition and Domain Knowledge](#_Problem_Definition_and)
   1. Problem Definition
   2. Domain Knowledge
      1. Literature Review
      2. Test Open-Source Models?
3. [Data Collection and Sourcing](#_Data_Collection_and)
4. [Data Cleaning and Processing](#_Data_Cleaning_and)
5. [Exploratory Data Analysis](#_Exploratory_Data_Analysis)
6. [Model Building and Evaluation](#_Model_Building_and)
7. [Model Results](#_Model_Results)
8. [Model Deployment](#_Model_Deployment)

## Data Science Lifecycle Research/Review

A diagram of a process

Description automatically generated

## Problem Definition and Domain Knowledge

### Problem Definition

1. To forecast, with the highest possible accuracy, when PM 2.5 levels will be elevated in the Connecticut River Valley, specifically over the City of Keene, given meteorological data gathered by KSC’s Nora Traviss.
2. To evaluate each model to optimize for a variety of factors: error, compute, number of features, and more to be determined.

### Domain Knowledge

Ad Hoc Links-

<https://library.wmo.int/viewer/60113/download?file=100_2023-edition_en.pdf&type=pdf&navigator=1>

**Time Series Analysis**

<http://mayoral.iae-csic.org/timeseries2021/hamilton.pdf>

Preface / text navigation-

Chapter 13 on the Kalman filter could be covered after chapter 4, after chapter 12, or skipped altogether

Chapter 6 on Spectal analysis could be covered at any point after chpt 1 or skipped altogether

Although the book was written with an econometrics course in mind it should be useful in many other domains. It claims that “time series econometrics” has become almost synonymous with “empirical macroeconomics”

Chapter 1: Difference Equations

A difference equation is an equation relating an expression relating a variable y to it’s previous values.

This is a linear first-order difference equation:



This is first order because the first lag of the variable (yt-1) is used. It is linear because it takes the form y= mx+b

Goldfeld's (1973) estimated money demand function for the United States.



mt := log of real money holdings of the public

It := log of aggregate real income

rbt := log of the interest rate on bank accounts

rct := log of the interest rate on commercial paper

A black text with black text

Description automatically generated with medium confidence

When analyzing dynamic systems, we may simplify the other inputs to a scalar w value.

Q: If a dynamic system is described by [1.1.1], what are the effects on y of changes in the value of w?

Solving a difference equation by recursive substitution

A math equations on a white background

Description automatically generated

A math equations and symbols

Description automatically generated with medium confidence

Since we now know y1 and if we know w2,

A math equations on a white background

Description automatically generated

Dynamic multipliers:

1.1.7 expresses yt as a linear function, given the initial value y-1 and the values of w. To calculate the effect of w0 on yt. If w0 were to change with the same initial values, the effect on y would be:

A white background with black and white clouds

Description automatically generated

Remember that the partial derivative could be read as the rate of change of yt with respect to w0 equals theta to the t

TODO: I should try to do this derivation myself

A math equations and formulas

Description automatically generated with medium confidence

A math equations on a white background

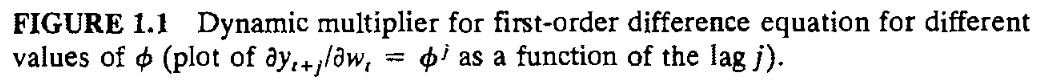
Description automatically generated

Note above that in the above calculation the partial derivative of wt is used to substitute theta to the second power in. Then you just need to address the term used for cancelling. It reads out as the rate of change of wt with respect to current income.

TODO: figure out why the cancelling term is equal to theta.

A group of graphs with numbers

Description automatically generated with medium confidence



As seen in figure 1.1a if 0<ϴ<1 then the multiplier  decays geometrically

Geometrically = Exponentially?

As seen in figure 1.1b if -1<ϴ<0 then the multiplier  decays alternating its sign

As seen in figure 1.1c if ϴ>1 then the multiplier  increases exponentially

As seen in figure 1.1d if ϴ<-1 then the multiplier  increases exponentially alternating sign

This reminds me of convergent, divergent, and oscillating series?

TODO: check this in undergrad notes

If ϴ=1 then the solution [1.1.9] becomes,

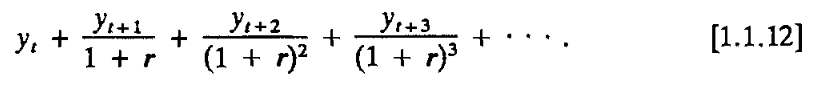


Since the theta coefficient is gone a one unit increase in w results in a one unit increase in y

A black text on a white background

Description automatically generated

If we were looking to find the effect of a change in w on a given value in the stream of future realizations of y, given a constant interest rate r>0. The present value at time t is:



A math equations on a white background

Description automatically generated

\*\*\* I don’t think this textbook is giving me enough valuable information for how obfuscated the language is. I think I should figure out what time series topics are relevant to weather/pm forecasting, smoothing, etc. then make sure whatever text I’m using has those in universally used terminology.

<https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm>

This is an engineering statistics handbook but I see some similarities with concept talked about in the video noted on below.

6.4 Intro

This text defines a “time series” as follows:

An ordered sequence of values of a variable at equally spaced time intervals.

The equally spaced time interval part is notable because the weather data I have is not equally spaced.

The fitting models listed in the introduction are:

 [Box-Jenkins ARIMA models](https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc44.htm)

 [Box-Jenkins Multivariate Models](https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc45.htm)

 [Holt-Winters Exponential Smoothing (single, double, triple)](https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc43.htm)

It states that it is beyond the scope of the handbook to covers all these. The overview will start with:

Averaging Methods and Exponential Smoothing Techniques

6.4.2 What are Moving Average or Smoothing Techniques?

Random variation is inherent to data taken over time. Smoothing aims to reduce or negate the effects of this random variation. When done properly this can more clearly show the trends and cycles of the data.

The two groups of smoothing methods are:

Averaging Methods and Exponential Smoothing Methods

The most basic averaging method is the simple ‘average’ of all past data points.

This will automatically produce the least mean squared error (MSE) of all past data points.

Examples given show that perhaps a simple average would be a good estimator for the amount of supplies delivered by a random given delivery driver based of the averages of other delivery drivers. They show it to minimize MSE.

This would however not be a good estimator for the profits of a computer technology company’s profits that show an upward trend in profit.

In general we can state that the simple average is only a good predictor if there are no trends.

We note that the average “weights” all past data points equally since the coefficient multiplied by each data point is 1/n. Then obviously these are summed.

A math equations on a yellow background

Description automatically generated

6.4.2.1 Single Moving Average

One method of smoothing may be to group successive datapoints and take their average.

A paper with numbers and text

Description automatically generated

6.4.2.2 Centered Moving Average

For the last example table. They placed the MA at the third interval according to the previous three values. If it was a centered moving average it would be start at the second interval and go until interval n-1. This works fine for odd smoothing time periods but what if the number was even?

If M=4 the centered MA’s would need to be placed at 2.5, 3.5, …

To avoid confusion we do another step of centered smoothing to place the values at 3,4,5,… instead

A yellow paper with black text and numbers

Description automatically generated

<https://www.youtube.com/watch?v=R4tcKNJe3xw>

Time series – collection of datapoint ordered in time

Properties –

* Stationarity

Mean and variance don’t change over time

A graph of a graph

Description automatically generated with medium confidence

* Periodicity

Seasonality or repeated fluctuations over time

A graph showing the flu cases

Description automatically generated with medium confidence

* Autocorrelation

Correlation of a time series with a delayed copy of itself

Ex. Is that temperature is greatly correlated with temperature from the previous day

Models for classification and prediction-

* Moving Average (MA)

A math equations and numbers

Description automatically generated with medium confidence

* Auto regression (AR)

A math equations and formulas

Description automatically generated with medium confidence

Combination of these two models is called an ARMA(p,q) model: Works well for stationary, low periodicity and autocorrelation

A math equations and numbers

Description automatically generated with medium confidence

Search: Time series analysis weather forecasting

<https://keras.io/examples/timeseries/timeseries_weather_forecasting/>

Search: Time series pm 2.5

<https://ieeexplore.ieee.org/document/9359734>

<https://www.mdpi.com/2073-4433/14/2/340>

<https://www.frontiersin.org/articles/10.3389/fenvs.2022.945628/full>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9800338/>

**Weather Modeling**

Search: weather modeling

<https://www.noaa.gov/jetstream/upper-air-charts/weather-models>

<https://www.weather.gov/rnk/models>

<https://www.mmm.ucar.edu/models/wrf>

<https://www.ncei.noaa.gov/products/weather-climate-models/numerical-weather-prediction>

<https://sitn.hms.harvard.edu/flash/2024/ai_weather_forecasting/>

Search: intro to weather forecasting

<https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.weather.gov/media/shv/education_resource_library/forecasting/Forecasting.ppt&ved=2ahUKEwiNvMe8qcOFAxUuEGIAHRSYAiYQFnoECB8QAQ&usg=AOvVaw1ZBYXN1FCxU54BWZG3_1Ib>

<https://svante.mit.edu/~jscott/12.310/IAP_2022_lecture1_JRS_topost.pdf>

**Particulate Matter Forecasting**

Search: particulate matter forecasting

<https://www.sciencedirect.com/science/article/pii/S1877050920312060>

<https://www.arl.noaa.gov/research/surface-atmosphere-exchange-home/o3-and-pm-2/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9723408/>

<https://www.mdpi.com/2073-4433/13/9/1451>

<https://ieeexplore.ieee.org/document/9359734>

There are dozens of papers on this topic which bodes well for an ML approach.

**Air Inversion Forecasting**

Search: air inversion forecasting

<https://www.weather.gov/media/lzk/inversion101.pdf>

<https://www.weather.gov/source/zhu/ZHU_Training_Page/Miscellaneous/inversion/inversion.html>

<https://www.alleghenycounty.us/files/assets/county/v/1/government/health/documents/air-quality/sadar-emplus-article-reprint.pdf>

<https://climate.usu.edu/inversion.php>

<https://pubs.aip.org/physicstoday/article-abstract/71/10/74/948000/Waking-up-to-temperature-inversionsWhen-cool-air?redirectedFrom=fulltext>

<https://link.springer.com/article/10.1007/s44274-023-00018-w>

## Data Collection and Sourcing

Data set is collected from a meteorological station on Water Street in Keene NH. Data was gathered and distributed by Dr. Nora Traviss for the purpose of forecasting PM 2.5 concentration in the Keene area.

TODO: Verify this information and contact Dr. Traviss or Dr. McGregor for updated data.

## Data Cleaning and Processing

Table 1: Data Features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Code** | **Column** | **Descriptive Name** | **Type** | **Nulls** |
| F0 | Datapoint Number | Row Number | Int [0,+] | No |
| F1 | Datetime | Datetime Object | Datetime | No |
| F2 | Created At | Text Date and Time | Date + Time |  |
| F3 | PM25 | Particulate Matter <2.5 Microns Diameter | Float [0,+] | No |
| F4 | Date | Date with less frequently updated time | Datetime |  |
| F5 | temp | Temperature in Fahrenheit | Float [-,100] | No |
| F6 | dewpoint | Dew Point | Float [-,+] | No |
| F7 | RH | Relative Humidity | Float [0,100] | No |
| F8 | windDIR | Wind Direction in Degrees | Float [0,360] | No |
| F9 | windMPH | Wind Speed | Float [0,+] | No |
| F10 | precip | Precipitation | Float [0,+] | No |
| F11 | mslp | Mean Sea Level Pressure | Float [0,+] | No |
| F12 | visibility | Visibility | Float [0,10] | No |
| F13 | gust | Gust Speed | Float [0,+] | No |
| F14 | wxcodes | Weather Codes | String | Yes |
| F15 | (top) Snow Depth (in) | Top Snow Depth | Float [0,+] | No |
| F16 | (middle) Snow Depth (in) | Middle Snow Depth | Float [0,+] | No |
| F17 | (bottom) Snow Depth (in) | Bottom Snow Depth | Float [0,+] | No |
| F18 | (top) Snow Temp. (deg. F) | Top Snow Temperature | Float [0,+] | No |
| F19 | (middle) Snow Temp. (deg. F) | Middle Snow Temperature | Float [0,+] | No |
| F20 | (bottom) Snow Temp. (deg. F) | Bottom Snow Temperature | Float [0,+] | No |
| F21 | (top) Snow Density (%) | Top Snow Density | Float [0,+] | No |
| F22 | (middle) Snow Density (%) | Middle Snow Density | Float [0,+] | No |
| F23 | (bottom) Snow Density (%) | Bottom Snow Density | Float [0,+] | No |
| F24 | Date w/o Time | Date | Date | Yes |
| F25 | Hour | Hour | Float [0,24] |  |
| F26 | Forecasted from 0 UTC | Forecasted from 0 UTC | String | Yes |
| F27 | FEW | Few Cloud Layer | Int [0,12000] | Yes |
| F28 | SCT | Scattered Cloud Layer | Int [0,12000] | Yes |
| F29 | BKN | Broken Cloud Layer | Int [0,12000] | Yes |
| F30 | OVC | Overcast Cloud Layer | Int [0,12000] | Yes |
| F31 | VV | Vertical Visibility | Int [0,12000] | Yes |
| F32 | Clouds | Cloud Coverage | [0,8] | Yes |
| F33 | Clds1000 | Cloud Coverage at 1000 ft | [0,8] | Yes |
| F34 | Clds2000 | Cloud Coverage at 2000 ft | [0,8] | Yes |
| F35 | Clds3000 | Cloud Coverage at 3000 ft | [0,8] | Yes |
| F36 | Clds4000 | Cloud Coverage at 4000 ft | [0,8] | Yes |
| F37 | Clds5000 | Cloud Coverage at 5000 ft | [0,8] | Yes |
| F37 | Clds6000 | Cloud Coverage at 6000 ft | [0,8] | Yes |
| F38 | Clds7000 | Cloud Coverage at 7000 ft | [0,8] | Yes |
| F39 | Clds8000 | Cloud Coverage at 8000 ft | [0,8] | Yes |
| F40 | Clds9000 | Cloud Coverage at 9000 ft | [0,8] | Yes |
| F41 | Clds10000 | Cloud Coverage at 10000 ft | [0,8] | Yes |

### Data Cleaning

Replace missing values

Which features need this: WX Code, Clouds

Could I make the data more usable by regularizing the time delta between data readings. Perhaps just duplicate rows of data to more evenly space out the time of recording.

Remove erroneous duplicate values?

Ex. Dewpoint and RH

Enumerate categorical values?

Ex. WX Codes

Address data imbalances

Ex. Lack of high pm 2.5 instances

### Feature Engineering

Links-

<https://www.heavy.ai/technical-glossary/feature-engineering>

This one says that the four processes of feature engineering are: Feature Creation, Transformations, Feature Extraction, and Feature Selection.

<https://www.kaggle.com/code/prashant111/a-reference-guide-to-feature-engineering-methods>

<https://www.geeksforgeeks.org/what-is-feature-engineering/>

<https://www.mathworks.com/videos/applied-machine-learning-part-1-feature-engineering-1547849284703.html?gclid=CjwKCAjwoPOwBhAeEiwAJuXRhxOuu7ViZoIU1n_A9-xTUaWPYpjW8zflNAv1KpgXnjZHAQ8nPQLhsRoCrnAQAvD_BwE&ef_id=CjwKCAjwoPOwBhAeEiwAJuXRhxOuu7ViZoIU1n_A9-xTUaWPYpjW8zflNAv1KpgXnjZHAQ8nPQLhsRoCrnAQAvD_BwE:G:s&s_kwcid=AL!8664!3!678713281811!p!!g!!feature%20engineering&s_eid=psn_155396784615&q=feature+engineering&gad_source=1>

Data augmentation-

Each feature \* Each feature

## Exploratory Data Analysis

<https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf>

Summary stats

Histograms, boxplots, scatterplots

Correlation analysis

Hypothesis testing

## Model Building and Evaluation

## Model Results

## Model Deployment