STAT 773 Final Project

Land Temperatures Over Time and Global Warming

Edward Rao

Dec 2021

**Intro and Data Outline:**

In the age of globalization and increased connectivity it’s very easy to access information of all sorts from across the world in seconds. While this has led to a faster and more efficient way for all of mankind to connect to our fellow human beings, it has also led to the rise of misinformation, distrust, and polarization. One of the biggest “fake news” stories continually perpetuated by many people is the idea of global warming and climate change being fabricated as a hoax, whether by the government or scientists is up to interpretation. In this project we will look at the changes in average global temperatures over time, observe patterns, and make predictions by using statistical time series models that we’ve learned over the course of this class.

Our dataset is a csv file obtained from data.world and consists of several columns, but many of these other variables are empty due to the lack of reliable data from earlier dates in the 18th and 19th centuries. These values will not be useful to our study (as they play a role in other subset datasets for particular countries, which are more detailed and have time-period specific information), but are factors to keep in mind, particularly the temperature uncertainties and the contrasting levels between land and sea temperatures.

Our main predictor and response variables will be the dt by month (delta time, we will refer to it as time) and average land temperature across the globe, measured in degrees Celsius. Because the variable tracks global temperatures, including areas in Northern Canada and Siberia along the North Pole and Antarctica in the South, the measures appear to be a bit lower than what we would expect. This, however, should not affect the series over time since all the values are consistent and measured in the same manner.

The other variable that we will not be using but contains data is the uncertainty factor for the temperature. It is measured by the 95% confidence interval around the average global temperature. While it won’t play a role in our models, it’s something to keep in mind when looking at the results for analysis.

For subsequent analysis we will be creating month and year indicator variables for the dataset, too.



Graphical user interface, application, table

Description automatically generated

Our time variable ranges from 1750 to 2015. However, there are missing data points in the first 3 years, which can lead to incomplete readings in the data (and that would cause Minitab to freak out/not work properly). Thus, instead of having to perform operations on the data or change the whole dataset to account for these few missing points, these years were simply removed. Instead, the data begins from 1753, and these initial 3 years do not detract from the analysis in the grand scope of things. There are a total of 3156 data entries, a total of 263 years split by month.

**Question and Hypothesis:**

The question we are exploring is whether a trend in this seasonal data could be detected and what those trends could mean. In the case of our dataset, we want to answer the question of whether or not there is a change in average global temperatures over time and what this could mean.

For global temperatures, we would expect to see the average land temperatures rise whether gradually or suddenly, particularly in more recent years starting from the Industrial Revolution all the way until present day and global warming. Temperatures should also drop on average in the winter and be higher during the summer due to the Earth’s tilted axis and distance from the Sun. Our model should fit these patterns and we would expect any predictions we make to follow them.



**Initial Data Analysis:**

Chart, scatter chart

Description automatically generated

It’s somewhat difficult to make any predictions at a glance, particularly with so many data points (A small section of the initial 7 years is located in the Appendix simply for clarity, with ACF to show the similar nature). There does seem to be a slight shift upwards in the trend in the later entries (aka more recent years), and aside from the initial few hundred points the values also have far less variation in the later years.

If we make an initial trend analysis plot, it looks like this:

Chart

Description automatically generated

And the quadratic option:

Chart

Description automatically generated

Both plots have a close fit, but there does seem to be some change, however slight. Due to negative values the exponential plot is not included. And the quadratic plot cannot actually be used because some of the temperature values are sub-zero, which cannot have a square root taken of the value. Either way it’s close enough to not make a difference, particularly as it mainly plays a role in regression, which we will soon see the shortcomings of in the modelling and testing section for this particular dataset.

If there was any doubt to the seasonal patterns in the data, here is the ACF for the temperature.

Chart

Description automatically generated

With a seasonal “repeat” of approximately 12 lags (aka per 12 months), our results make sense with our data, since seasonal temperatures should be expected to vary.

**Modelling and Fitting:**

We now have a plethora of options available to us for fitting an appropriate model to the data. Fortunately, the data is very suited for regression, trend and seasonal models, decomposition, smoothing, and autogregressive integrated moving averages (ARIMA) techniques alike, since there is a clear seasonal aspect that repeats throughout the years that can be handled and outliers/errors APPEAR (for now) to be few and far between, making it easier when it comes to making our predictions and forecasts.

Below is a cursory list of some models that can be used.

Linear regression model (unlikely)

Linear regression model with categorical month variable

Linear regression model via sine-cosine trigonometry

Decomposition model

Single exponential smoothing model (unlikely)

Double exponential smoothing model (unlikely)

Holt-Winters model

ARIMA model

While a no-trend regression model can also be created and used, because of the seasonal component it would be more helpful to establish predictions that line up with these variations.

For our predictions we can predict the average global temperature for years after 2015 and check with the real recorded temperatures online via the same source this dataset used (Berkeley Earth, link in References). The data is as follows for reference:

Table

Description automatically generated

We can just use 2016-2017 as comparison and validation. Note the above values are not the values themselves but instead “anomalies”, which means they are compared to the values below.

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

2.60 3.21 5.30 8.30 11.30 13.44 14.31 13.84 12.05 9.21 6.08 3.63

So if we were to translate the above values we would get:

2016: Table

Description automatically generated

2017: Table

Description automatically generated

We will begin with regression, which can be applied in several different ways. Using the linear trend, we can set time as the continuous predictor (via an equivalent index variable) and see the following:

Table

Description automatically generatedText

Description automatically generated with medium confidence

Graphical user interface, text

Description automatically generated

While the value is overall significant (with p-values less than .05), our r-squared percentage of 0.67% is rather abysmal to be generous. Our Durbon-Watson statistic also does not inspire confidence in this model.

Chart, line chart

Description automatically generated

And obviously the model is not normally distributed with such a low p-value. This should be a common occurrence because of the trend and seasonality combined, which we will see in the other regression models as well.

Linear with month as categorical:

Here, we add the month indicator variable as a categorical variable, which will allow month to play a role. The differences are very clear:

Table

Description automatically generated

Table

Description automatically generated with low confidence

Text

Description automatically generated with medium confidence

Comparatively, this model is much better when it comes to the r-squared value and DW statistic. While the latter implies there are still some autocorrelated errors present, likely from the initial data with greater variation, likely due to more unreliable data collecting methods in the 18th and early 19th centuries.

Chart, scatter chart

Description automatically generated

Testing for normality, we see the following.

Chart, line chart

Description automatically generated

So unfortunately due to the low p-value the model is also not normally distributed. In order to account for autocorrelation we can take a lag of order 1, aka shifting all our values back by 1 month and see if it fixes the issue.

Table

Description automatically generated Graphical user interface, text

Description automatically generated

Our p-values and r-squared value remain at good values, but the DW statistic has improved significantly. Looking at the fits graph there are also very few errors out of bounds.

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

Because of the lack of normality and similar shape in all of the normality plots, it should be reasonable to conclude from the graph that the distribution of the errors in the base temperatures plot is irregular due to the presence of a change in trend over time. More specifically, the values in recent years are higher and fall out of our 95% confidence bounds for future predictions, and implying that there are outliers in the data’s temperatures during some years. Also, recorded data in earlier years is likely less reliable due to the practices and equipment of a pre-Industrial Revolution era.

We can also use the trigonometric method for regression and see if there are any substantial differences present.

Linear with trigonometry:

First, we perform a test of equal variances to check if the standard deviations for all the years are the same. If they are it means constant seasonal variability, which means that there is a seasonal trend in the data, which then means patterns exist that can be approached via trigonometric means (and in general is a way to also test for seasonality as well).

Chart

Description automatically generated

With a p-value of 1 our test shows a very clear seasonal pattern occurring every year, if not evinced by the previous data analysis.

Sin2 Cos2

Since seasonal trends can be equated as regular, predictable change over an interval, we can use a sine-cosine trigonometric function to help fit our model. We can apply sincos2 (or a half period) to our time continuous variable and make a prediction that would fit to such a pattern.

Table

Description automatically generated Graphical user interface, text

Description automatically generated

The performance is very similar to our linear regression model using month as a categorical variable. This makes snese mathematically, since we observed earlier the seasonal trend that both the month and trigonometric representation of seasonal (monthly) values would reflect. The DW statistic is, again, somewhat concerning. A lag would yield better results:

Table

Description automatically generated Graphical user interface, text, application

Description automatically generated

Like in Exam 1 and in class, the results are very similar to the categorical predictor version of the model, likely due to the data we’re working with. Including, unfortunately, the normality test, too, which still has a p-value of under .05:

Chart, line chart

Description automatically generated

Sin4Cos4 were also used, but performed considerably worse, with low DW and r squared values (Appendix).

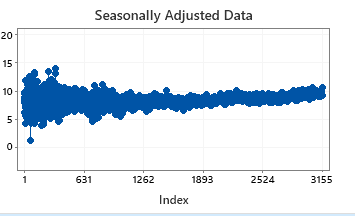
Decomposition:

With our large number of observations and the presence of both trend and seasonal components we can also use decomposition methods to make forecasts for our data. As decomposition aims to separate the individual components of our series, it will allow us to use this model to accurately track our parameters. We will be using the additive model due to no change in parameters over time, but the multiplicative model is included without commentary in the Appendix for visual reference and comparison.

Chart

Description automatically generated

While not a perfect fit, it’s certainly more additive in nature than multiplicative. The margins of error in the beginning are supported when we look at the seasonally adjusted data:



Some useful information can also be gleaned via looking at the residuals, where we see that during the spring and summer seasons the residuals tend to shift upwards, signifying the rising temperatures during, but not limited to, these months.

Chart, box and whisker chart

Description automatically generated

Exponential Smoothing:

As we’ve established our dataset has trend and seasonal components, only one exponential smoothing model really fits for our time series, the Holt-Winters Method exponential smoothing model. However, unlike quadratic or exponential regression, it is still possible to visualize the single and double smoothing models, although much of the variation will not be accounted for. These will be located along with the others in the Appendix.

As for Holt-Winters Method, we can compare the performance of the difference in values between the weights for level, trend, and seasonal components alike can affect the outcome of our predictions. For the initial values all set at 0.2, we see the following results:

Chart, histogram

Description automatically generated

As an addendum, even though we’ve established the additive nature of the trend over multiplicative, the multiplicative method for Holt Winters can seal any doubts we would have otherwise:

Chart, histogram

Description automatically generated

Not even close.

By changing the values all to 0.4 we can see that the higher emphasis on current values actually makes the smoothing a lot worse. Even tinkering with the values of the constants do not yield a better model than setting them to 0.2 across the board.

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generated

Ultimately, we’ll stick with the 0.2 across the board.

ARIMA:

We can first take a look at the PACF function to determine the order the non-seasonal difference should be taken on.

Chart

Description automatically generated

First order difference non-seasonal:

Chart

Description automatically generated

Actually just a flat line, so very good. We can establish for our ARIMA function a 1st order difference non-seasonal for the “I” component.

ACF of non-seasonal difference lag 1 data:

Chart

Description automatically generated

MA value 1, almost certainly; next is PACF.

Chart

Description automatically generated

There’s a lot of noise at the beginning, which fits with our metadata since we know that the early data collected is not as reliable (and checking the data’s uncertainty values confirms this).

We can go with a value of 1 or 2. Anything more could work, though, but we should not use high values if low values give us a good result to needlessly complicate the model.

Next to get the seasonal variables we look at the detrended, but non-stationary data.

Chart, scatter chart

Description automatically generated

As it’s seasonal we must take a seasonal difference to get our values (as demonstrated by the ACF)

Chart, box and whisker chart

Description automatically generated

The seasonal difference is of lag 12, which makes sense due to the series being monthly, as mentioned before. The ACF then shows

Chart

Description automatically generated

There’s maybe some sort of seasonality present, but not apparent. We can set the MA value to 1 or 2. Next, the PACF:

Chart

Description automatically generated

And then the AR value can be set to 1, making our ARIMA model:

ARIMA(1/2/3/etc,1,1)x(1,1,1/2) or

But Minitab is not ok with a lot of the models, and other don’t perform so well; we use the choice of (3,1,1)x(1,1,1) to demonstrate the best result.

Table

Description automatically generated Text

Description automatically generated

The p-values look good, the MS is low, and it passes the chi-square test. Looks good to me!

**Forecasts and Performance Comparison**

Linear regression model with categorical month variable

By using Minitab’s predict alongside the lag values of the actual results listed above in the Modelling and Fitting section, we obtain the following predictions and bounds for the model:

Text

Description automatically generated

Table

Description automatically generated

When compared with the originals, the graph looks like this:

Chart, line chart

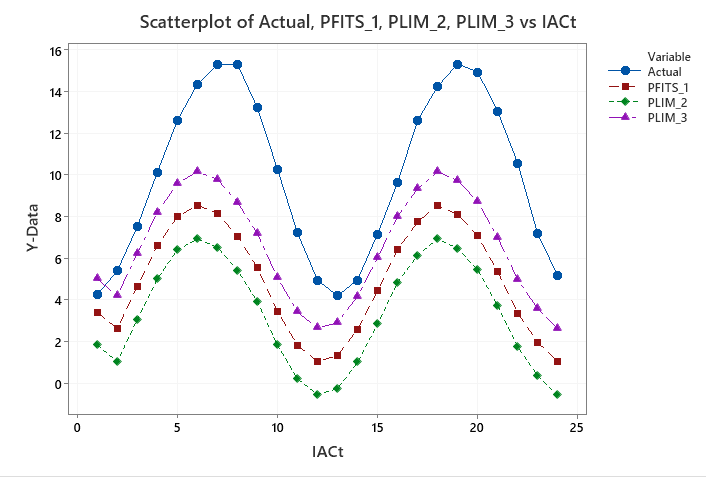
Description automatically generated

Linear regression model via sine-cosine trigonometry Text

Description automatically generated

Table

Description automatically generated



The plot looks almost identical to the linear model with month, and although the values are slightly different the overall shape and predictions points for the forecasts are generally very much the same.

Decomposition model

Chart, line chart

Description automatically generated

Sadly Minitab does not provide an option to store upper and lower bounds when generating forecasts. However if we use the formula for calculating the SDs provided by Minitab themselves (References) we can use:

A screenshot of a computer

Description automatically generated with medium confidence

Where 1.96 includes the 95% confidence interval and 1.25 an “approximate proportionality constant” that when multiplied by MAD (Mean Absolute Deviation) aka

Diagram

Description automatically generated

We can actually set the boundaries ourselves!

We calculate using the formula the value of MAD

MAD = sum(Actual-forecast)/24 = 0.981561

We then plug it into the formula to get our bounds:

Upper limit = Forecast + 1.96x1.25x0.981561

Lower limit = Forecast – 1.96x1.25x0.981561

Chart, line chart

Description automatically generated

Now that’s more like it!

Holt-Winters model

Chart, line chart

Description automatically generated

ARIMA model

Chart, line chart

Description automatically generated

Summary, Conclusions, Future Research

Our results show that many of the linear regression models, while passing most of the tests and limitations we covered this semester (save for normality -\_-), were not robust enough to handle the series, in particular its trend behavior.

The other methods, however, were all very close. Holt-Winters and ARIMA especially were very close, and it’s difficult to say which one is better, as ARIMA’s fits are much closer than Smoothing, but Winters’ projections have a narrower prediction range which can be better in the long-term when it comes to future predictions. Due to the context of the dataset I’d have to give it to Holt-Winters’ method, but perhaps there’s an ARIMA combination that could have yielded even better results.

When it comes to this dataset, looking back it’s very clear that temperature changes are very seasonal in nature, and that a lot of the difficulty regression may have had with it had to do with the outliers and more uncertain data that was recorded from the beginning in earlier years. Regardless, the pattern that we observe is apparent. If we take a look at the early data (with all its inconsistency and variability) vs the recent data we can see the change.

Chart, line chart

Description automatically generated

There is very clearly a rise in temperatures, particularly in the winter months. These changes may not seem as “severe” as the graph implies, but these measures are of temperatures on a global scale, and an increase in a single degree Celsius globally could spell dire consequences not just for animals and natural habitats, but humanity as well.

When it comes to the future and climate change, it’s difficult to provide an easy solution. As statisticians while it may not be our job to address the issue head-on (whether it be via renewable energy or reduction in fossil fuel usage, idk) we are obliged to present the data and let the conclusions speak for themselves.

For the future of global warming, all we can do is create models, draw conclusions, and hope the results can inspire some change. With how busy and interconnected everything it, it’s easy to get lost and swept up in current events like the ongoing pandemic, politics, or personal life, but sometimes looking at the facts can help provide us some perspective with the world around us and hopefully give some certainty in these uncertain times.

Appendix:

A picture containing text, indoor

Description automatically generatedChart

Description automatically generated

SinCos4

Table

Description automatically generated

Text

Description automatically generated

Multiplicative Decomposition Model:

Chart

Description automatically generated

Single EXP Smoothing

Chart

Description automatically generated

Double Smoothing (with trend):

Chart

Description automatically generated

References:

<http://berkeleyearth.org/data/>

<https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/time-series/how-to/single-exponential-smoothing/methods-and-formulas/methods-and-formulas/>

Thanks for reading, and thanks for the semester! Even though I’m not a STATs student, I definitely learned a lot in this class that will help me in the future. Hope you have a Happy Holidays!