Spring or Shitsville? A statistical analysis of alternative season classifications

Andrea Knox

17 January, 2023

*This article is based on work done by Dr Andrea Knox and Jordan Monk in fulfillment of the group project requirements for the Victoria University of Wellington course: STAT394 Multivariate Statistics. You can find the full report, data, scripts, and reports in our [Github repository](https://github.com/andreaknox-nz/real_seasons).*

### You can’t beat Wellington on a good day. Unfortunately you never know when those days will be

*“I wish I was in Wellington, the weather’s not so great.”*  
The Mutton Birds, “Wellington”, 1994.

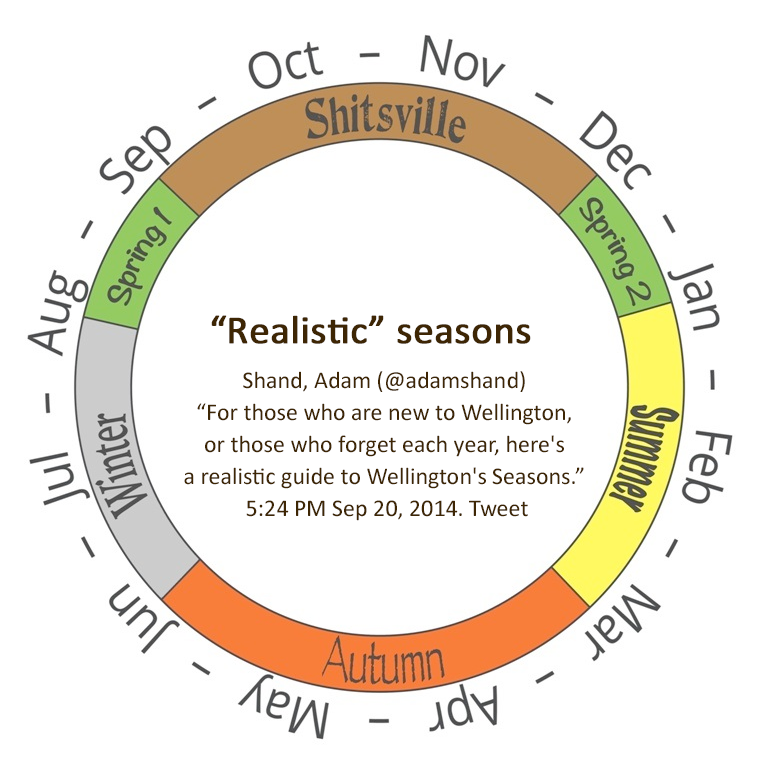
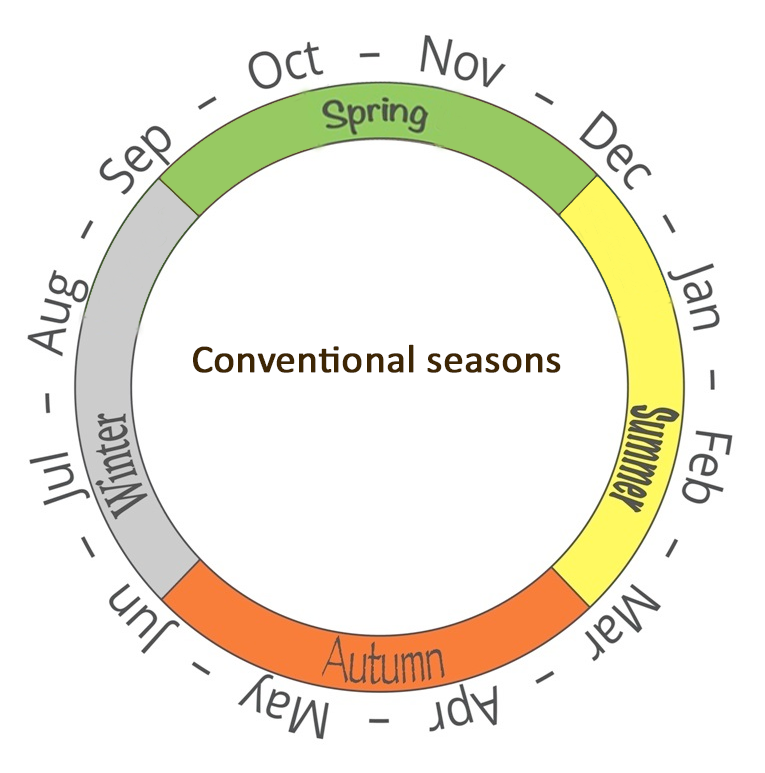
Wellington is notorious for its windy and unpredictable weather. We celebrate it in our public sculptures, our songs, and our (seemingly daily) complaints about it. Our weather seems to defy the seasons - we have hail storms in summer and gloriously sunny winter days. Spring can seem like a cruel joke. We envision hope and renewal, newly opened flowers and crisp sunny mornings. Instead we endure grey rainy days and a howling wind. Perhaps we need to adjust our expectations.



The ‘blown away’ sign in Miramar, Wellington. Image by Wainuiomartian, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

### We can’t change the weather but maybe we can change the seasons

In a popular 2014 [tweet](https://twitter.com/adamshand/status/513197000930521089?lang=en), Adam Shand proposed a new classification of Wellington’s seasons that splits spring into two periods (August and December) and renames what was formerly spring (September to November) as “Shitsville”. People approved. There have been hundreds of retweets, you can [buy the t-shirt](https://shitsville.printmighty.co.nz) and [there’s even a website](https://www.realnzweather.com) that tells you what the “real” season is now.



### Can we make an evidence-based case for official change?

So why not officially adopt these ‘real’ seasons? In a country where a [bat can win a bird-of-the-year competition](https://en.wikipedia.org/wiki/Bird_of_the_Year), and a [kiwi shooting lasers from its eyes](https://nzhistory.govt.nz/media/photo/fire-lazar) can almost become our national flag, redefining the seasons and calling one Shitsville doesn’t seem like too big a stretch!

But behind every great policy is ~~great~~ some evidence and as-yet no evidence for this proposal exists. Humans are notoriously prone to confirmation bias. Are those of us who are ‘real’ season true believers merely seeing the weather patterns we expect to see?Or do the ‘real’ season actually better describe our weather?

We decided to find out.

### We used weather data to compare the conventional and ‘real’ season classifications

We obtained five years of daily weather data from the National Institute of Water and Atmospheric Research (NIWA) and used four different statistical methods to compare how well the conventional and ‘real’ seasons classify actual weather patterns. We did this for Wellington and Auckland and got similar results. Here I only describe the Wellington results (to reduce length - you’re welcome) but you can find the Auckland analysis in our [technical report](https://github.com/andreaknox-nz/real_seasons/blob/main/final_report%20(technical)/Final_report.pdf).

Our weather data had five measurements taken every day during the five-year period 2017 to 2021:

* the maximum temperature recorded in the 24-hour period (in degrees Celsius).
* the minimum temperature recorded in the 24-hour period (in degrees Celsius).
* global radiation, which measures the total radiation from the sun during the 24 hours (in megajoules per square metre).
* wind run, which measures the distance travelled by surface wind during the 24 hours (in kilometres).
* the rainfall received during the 24 hours (in millimetres)

### A better classification puts days with similar weather into the same season and separates days with different weather into different seasons

Before we get to results, we need to talk about what makes a good classification. There’s always some subjectivity in judging what is ‘good’, but following [evaluation best practice](https://www.betterevaluation.org/frameworks-guides/rainbow-framework/frame/determine-what-success-looks), we can at least be transparent about our criteria.

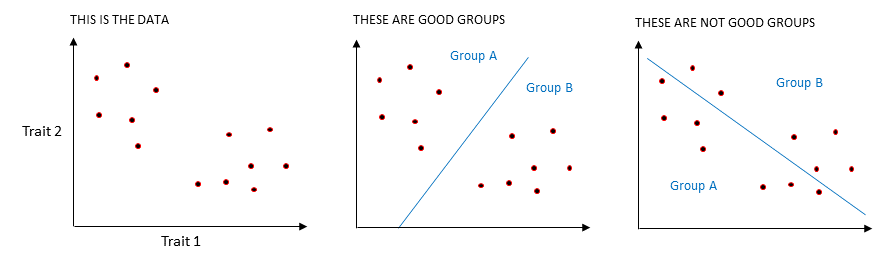
In this work, we say that a classification is good if it:

* puts similar things into the same group
* puts different things into different groups.

So one classification is better than another if:

* its within-group differences are smaller
* its group-to-group differences are larger.

The diagram below demonstrates this in terms of distance. Imagine that you are looking down on a room of people and each red dot is a person. We see two clusters of people and intuitively we group them as shown in the middle chart. This is a good classification because the distances between people in the same group are small and the distances between people in different groups are larger. Alternatively, we could form two groups as shown in the right hand chart. This would be bad. Why? Because there are large distances between some people in the same group and short distances between some people in different groups. The classification in the middle is clearly superior.



Some classifications are better at grouping things than others.

Now imagine that the distances between people represent something else - say, eye colour (Trait 1 in the chart above) and hair colour (Trait 2). In one cluster people have varying shades of blonde hair and blue eyes and in the other they have shades of dark hair and brown eyes. The best classification separates the two clusters, grouping together people whose hair and eye colours are close and separating people with more distant eye and hair colours.

Finally, imagine that the dots are days of the year. And the distances between them represent aspects of the weather. For example, Trait 1 could be rainfall and Trait 2 could be maximum temperature. A good classification groups together days with similar temperatures and rainfall and separates days with dissimilar temperatures and rainfall. This is the basis for how we decide whether the conventional or ‘real’ seasons are better. The better classification will have, on average:

* shorter weather-based distances between days of the same season
* larger weather-based distances between days of different seasons.

All of the analysis we describe below uses this fundamental concept.

### Standard deviations suggest that the conventional seasons are better at grouping similar days together

The standard deviation measures variability. Specifically, it measures the average distance of each observation from the centre (or the mean) of the group. For example, say we have only two days, one with a maximum temperature of 10 degrees and the other with a maximum temperature of 24 degrees. The mean maximum temperature is . The standard deviation is the average distance of each day’s measurement from , so in this case, it is .

For our analysis, we computed the standard deviations for each weather measurement, by season. And then we compared the standard deviations of the conventional seasons with those of the ‘real’ seasons. Smaller standard deviations indicate shorter (weather-based) distances between days of the same season and therefore a better classification.

Standard deviations (SD) for the conventional seasons, Wellington

| Season | SD(max temp) | SD(min temp) | SD(global radiation) | SD(wind run) | SD(rainfall) |
| --- | --- | --- | --- | --- | --- |
| Summer | 2.85 | 2.62 | 7.55 | 174.05 | 8.10 |
| Autumn | 3.07 | 2.96 | 5.37 | 178.15 | 8.76 |
| Winter | 1.88 | 2.38 | 2.99 | 198.46 | 9.07 |
| Spring | 2.75 | 2.82 | 7.11 | 185.03 | 7.92 |
| Average across seasons | 2.64 | 2.69 | 5.75 | 183.92 | 8.46 |

Standard deviations (SD) for the ‘real’ seasons, Wellington

| Season | SD(max temp) | SD(min temp) | SD(global radiation) | SD(wind run) | SD(rainfall) |
| --- | --- | --- | --- | --- | --- |
| Summer | 2.82 | 2.67 | 7.21 | 171.22 | 6.33 |
| Autumn | 3.07 | 2.96 | 5.37 | 178.15 | 8.76 |
| Winter | 1.94 | 2.39 | 2.36 | 208.62 | 9.70 |
| Spring 1 and 2 | 3.99 | 3.66 | 9.08 | 179.37 | 9.24 |
| Shitsville | 2.75 | 2.82 | 7.11 | 185.03 | 7.92 |
| Average across seasons | 2.91 | 2.90 | 6.22 | 184.48 | 8.39 |

In the tables above, we see that the average standard deviations are smaller for the conventional seasons than they are for the ‘real’ seasons, for every weather measurement except rainfall. And when we compare the standard deviations of *spring 1 and 2* with conventional *spring*, we see that the *spring 1 and 2* standard deviations are larger and are, in fact, the largest of any of the seasons, for every measurement except wind run.

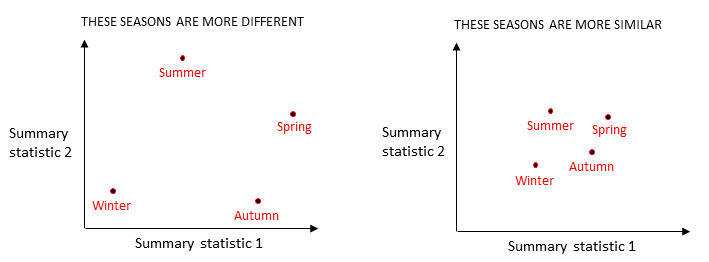
Overall, this suggests that conventional seasons may be better at grouping together days with similar weather and that *spring 1 and 2* may be grouping together days with dissimilar weather. So the conventional season classification looks better because, on average, it has shorter weather-based distances between days of the same season.

But what about our second criterion: a better classification has larger weather-based distances between days of different seasons. We looked at this next.

### Summary statistic-based distances between seasons suggest that conventional seasons are better at separating out different weather patterns

Whenever we group objects together we can use summary statistics to describe the overall characteristics of the group. Summary statistics are mostly quite straightforward. They are things like maximum values, minimum values, medians (the middle value), and means (described above).

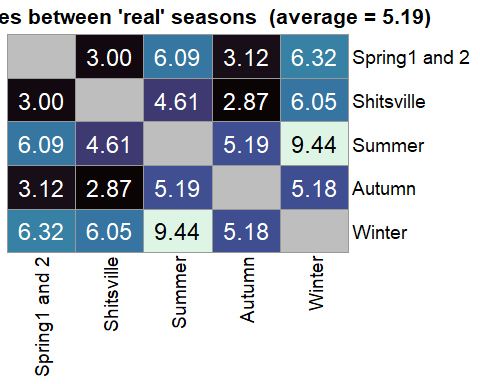
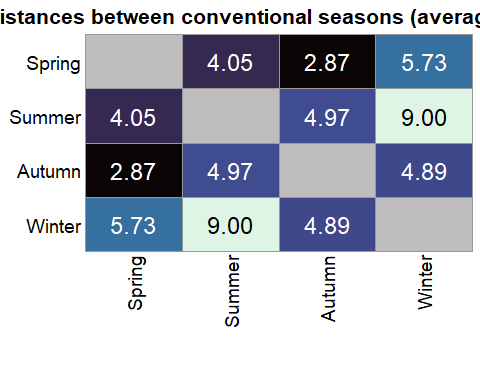
We can use each season’s weather data to compute summary statistics. And then we can use the values of those summary statistics to compute distances between the seasons. For example, in the charts below, imagine that summary statistic 1 is the maximum daily rainfall value recorded for each season and summary statistic 2 is the maximum wind run value recorded for each season. In the left hand chart these values are quite different and so the seasons are far apart. In the right hand chart the values are much more similar and so the seasons are closer together. The result in the left hand chart is better. It suggests that this classification is better because its seasons are more different to each other.



We can use distances based on summary statistics to see how similar seasons are to each other.

Similarly, we could use three summary statistics and then our distances between seasons would be computed as if they were in a three-dimensional space. Same concept. Easy. Now here’s where we break your brain. In fact, we used 30 different summary statistics. So we computed the distances between seasons as if they were in a 30-dimensional space. Can you visualise that? No? Me neither. This is unintuitive for us, living as we do, in a measly three-dimensional universe. By analogy though, it’s simple. We computed the distances based on 30 summary statistics exactly as we would in a two- or three-dimensional space - we just used more dimensions.

So what did we find out?

 In the charts above you can follow each grid across and down to find the distances between pairs of seasons. For example, the distance between *summer* and *winter* was 9.00 for the conventional seasons (left hand chart) and 9.44 for the ‘real’ seasons (right hand chart). And the distance between *spring* and *autumn* in the conventional seasons is the same as the distance between *shitsville* and *autumn* in the ‘real’ seasons because the ‘real’ season classification simply renames *spring* to *shitsville* and doesn’t change *autumn*.

The ‘real’ seasons do seem to be better at separating *winter* and *summer* from each other and from *spring* and *autumn*: all of these distances are longer between the ‘real’ seasons. This is perhaps not surprising given that *winter* and *summer* are both only two months long in the ‘real’ seasons (*spring 1 and 2* takes a month from *winter* and a month from *summer*). It may be easier for a shorter season to be more different to the others.

However, there are some quite short distances among the ‘real’ seasons, especially between *spring 1 and 2*, *shitsville*, and *autumn*. The average distance between seasons is, in fact, slightly shorter for the ‘real’ seasons (5.19) than for the conventional seasons (5.25).

And importantly, the distance between *spring 1 and 2* and *shitsville* is short: the second shortest of all the distances that we computed. This suggests that a key conjecture of the ‘real’ season classification: that *spring 1 and 2* and *shitsville* have different weather patterns, may not hold.

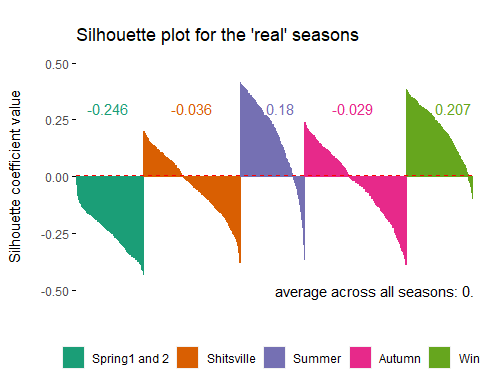
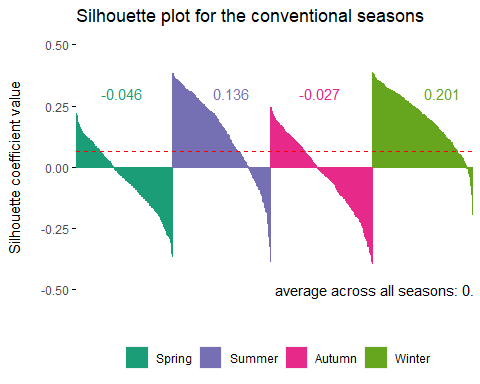
Overall, this suggests that the conventional seasons may be slightly better than the ‘real’ seasons at separating out days with different weather patterns,. But the conventional seasons aren’t much better. We would not call these results conclusive. Let’s try another approach!

### Silhouette plots show that conventional seasons are better at grouping days into appropriate seasons

We used silhouette plots to compare how good the conventional and ‘real’ seasons are at grouping days into the appropriate seasons. Silhouette plots use the silhouette coefficient, which is computed as described in our [technical report](https://github.com/andreaknox-nz/real_seasons/blob/main/final_report%20(technical)/Final_report.pdf). Go there if you want to see the maths. Less technically, how the coefficient works is as follows.

* It computes a representative point for each season, based on temperature, radiation, wind and rainfall measurements. In terms of distances, this point can be thought of as the middle of the season.
* It then uses each day’s weather data to calculate two distances: the distance from the day to the representative point of the season it is assigned to, and the distance from the day to the closest representative point of another season.
* It compares these distances. The resulting coefficient is between 0 and 1 if the day has been assigned to its closest season and between 0 and -1 if it has been assigned to a season that is not its closest.
* Higher coefficient values are better. The higher the value, the closer the day is to the ‘middle’ of the season it is assigned to.

You can then average the coefficients across days to get a measure of the extent to which days are appropriately classified into their closest seasons. Higher average values indicate a better classification.



In the silhouette plots above, the coloured wedges are made up of vertical lines: one for each day. Lines extending upwards from zero represent days that are classified into their closest season and lines extending downwards represent days that would be more appropriately put into a different season. It is bad news for the ‘real’ season classification that every day of *spring1 and 2* would be more appropriately grouped with a different season.

Looking at the averaged coefficient values (written in black under each chart and indicated by the dashed red line), we see that the average silhouette coefficient across all seasons is ten times higher for the conventional seasons (0.067) than for the ‘real’ seasons (0.006). This suggests that the conventional seasons are better at grouping together days with similar weather. Things are not looking good for the ‘real’ seasons!

But wait, there’s something else that we can see in these plots. In fact, neither the ‘real’ nor the conventional seasons do an especially good job. Both perform OK for *summer* and *winter*, with coefficients above zero for most days. But the other seasons: *spring*, *autumn*, and *shitsville*, are not at all good, with most coefficients below zero.

This brings us to our final analysis: what would an optimal weather-based classification look like and how do the conventional and ‘real’ seasons compare to that?

### K-means clustering can create better weather-based groupings of days

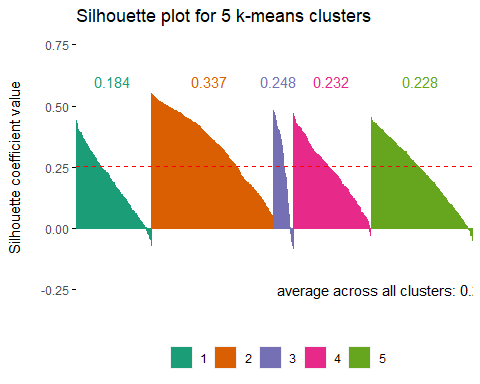
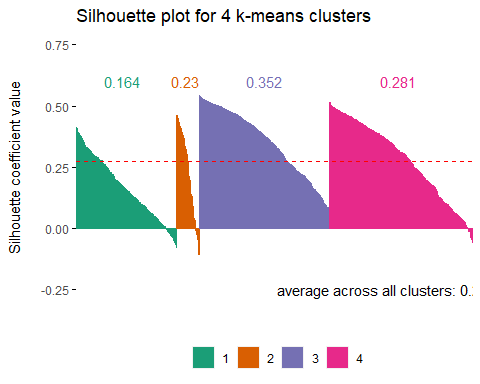
There are many statistical techniques for grouping objects together based on data. We used a method called ‘k-means clustering’ to group days into four or five clusters based on their weather.

In k-means clustering we specify how many clusters we want and then we use an algorithm to generate that number of clusters from the data. The algorithm works as follows.

1. The computer makes a random guess at where the centre of each cluster might be (this is not an informed guess, it is purely random).
2. The distance between each day and each centre is measured.
3. Each day is assigned to the cluster whose centre it is the closest to.
4. Each centre is then re-positioned to the middle of the cluster of days that were assigned to it.
5. The algorithm returns to step 2 and cycles through steps 2 to 4 repeatedly until the centres no longer move at step 4.

Here’s a [nice visualisation](https://www.youtube.com/watch?v=5I3Ei69I40s) of the algorithm in action.

We used this technique to generate four clusters (for comparison with the conventional seasons) and five clusters (for comparison with the ’real’ seasons). Since the clusters were generated from the weather data directly, they should be fairly optimal groups of days with similar weather patterns. Almost certainly they will be more optimal than the seasons. Were they? We can check using silhouette plots.



Yup, the charts above show that the average silhouette coefficients are consistently larger for the clusters than for the conventional and ‘real’ season classifications and that very few days have been grouped into inappropriate clusters.

### Conventional seasons are more similar than ‘real’ seasons to k-means clusters

Now that we have our more optimal groups, we can compare comventional and ‘real’ seasons to them and see which is closer to optimal. In this analysis we think of the k-means clusters as ‘correct’ and we test how good the seasons are at placing days into their ‘correct’ groups.

First we had to figure out how to match seasons to clusters. We want to give both classifications the best possible chance of generating correct predictions, so we tried all possible matches between seasons and groups and chose the concordance that returned the highest number of ‘correctly’ classified days. The best concordances were:

* conventional seasons matched to four clusters: *spring* = 1, *summer* = 4, *autumn* = 2, *winter* = 3.
* ‘real’ seasons matched to five clusters: *spring 1 and 2* = 3, *shitsville* = 1, *summer* = 5, *autumn* = 4, *winter* = 2.

We then computed four different performance metrics to compare how well the conventional and ‘real’ seasons make ‘correct’ predictions. For example in the conventional seasons, a *spring* day is correctly classified if it is in cluster 1, *summer* days are correct if they are in cluster 4, *autumn* in cluster 2 and *winter* in cluster 3.

The first of our four performance measures: overall accuracy, is straightforward. It is simply the overall percentage of correct predictions made by the classification. The higher the percentage the better. The other three performance measures are a bit more complicated and you can find a full description of them in our [technical report](https://github.com/andreaknox-nz/real_seasons/blob/main/final_report%20(technical)/Final_report.pdf). But all you really need to know is that bigger is better. The classification with higher percentages is more similar to the ‘optimal’ k-means clusters.

Performance of the conventional season classification against the four k-means clusters

|  |  |
| --- | --- |
| Overall accuracy | 48.3% |
| Macro Precision | 48.5% |
| Macro Recall | 43.0% |
| Macro F1 | 43.2% |

Performance of the ‘real’ season classification against the five k-means clusters

|  |  |
| --- | --- |
| Overall accuracy | 38.9% |
| Macro Precision | 40.2% |
| Macro Recall | 36.2% |
| Macro F1 | 35.9% |

On every performance measure, the conventional seasons did better than ‘real’ seasons.

We need to interpret this cautiously because performance here depends not just on how good the seasons are, but also on the how good the k-means clustering is. We might get a different result if we used a different clustering method. Nevertheless, the result is consistent with our other findings and provides one more piece of evidence that the conventional seasons are better than ‘real’ seasons.

However, consistent with what we saw in the silhouette plots, the performance measures suggest that even the conventional seasons are far from optimal, making ‘correct’ predictions less than half of the time.

### Unfortunately our evidence does not support officially changing to the ‘real’ seasons

Well, shit.

None of our findings support the idea that the ‘real’ seasons better describe Wellington’s weather (or Auckland’s either, see the [technical report](https://github.com/andreaknox-nz/real_seasons/blob/main/final_report%20(technical)/Final_report.pdf)). Instead of finding evidence to support an official change to the seasons, we found the opposite.

There are caveats with our results: we only used five years of weather data and one of our weather variables, global radiation (which measures total heat from the sun), is inherently tied to day length, which may give the conventional seasons an unfair advantage.

But perhaps we should remember something we learned in primary school: the seasons are, in fact, defined by the tilt of the earth relative to the sun. A proposal to redefine the seasons based on weather patterns might not get past a smart nine-year old’s bullshit detector.

### But all is not lost: we can keep Shitsville

Importantly though, it’s not the renaming of *spring* to *shitsville* that’s the problem. The culprit here is *spring 1 and 2*, which has very similar weather to *shitsville* and performs worse than conventional *spring* on the silhouette plots and the comparison of standard deviations.

So let’s ditch *spring 1 and 2* but continue to call *spring* *shitsville*.

And taking this further: in our analysis of distances, we saw that *shitsville* and *autumn* were the most similar of any pair of seasons. *autumn* could be considered to be a second *shitsville*, so why not rename it as such?

Here is my suggestion for a new, ‘more realistic’ re-envisioning of the seasons. Given the evidence, I believe it’s likely to be a better classification of the weather than the ‘real’ seasons, but further analysis is needed to confirm this. Unequivocally, however, it uses two more swear words than the ‘real’ seasons making it, therefore, objectively better.

