# Bike-Sharing Data Analysis: Prediction of Daily Bike Rental Counts Based on Multiple Linear Regression

Final Project Report · MA 575 Fall 2021 · C3 · Team #2

Ali Taqi, Hsin-Chang Lin, Huiru Yang, Ryan Mahoney, Yulin Li

12/10/2021

In this project, the following question is to be answered: If we have the past history of bike rental counts as well as records of environmental and seasonal conditions, how and how well could we predict the bike rental counts in the future? In this project, such questions are approached by predictive modeling of daily bike rental counts from a 2011-2012 Bike Sharing dataset [1]. The daily bike rental counts are predicted with models based on Multiple Linear Regression (MLS) using the environmental and seasonal variables as predictors. The initial goal of this project is to train the model using only the 2011 data, and then validate the prediction power of the model on the 2012 data. Given the limited time span of available training data, issues are found in the validation process using the 2012 data; the impact of user base on the future predictions is brought to our attention. The initial models are then revisited and corrected to account for the effect of user base. The refined models are expected to have better prediction powers than the initial MLS models, but a full validation would require further availability of bike rental data.

## 1 Introduction

Bike sharing has become a world-wide phenomenon. Optimization of inventories and dynamic reallocation of bike-sharing resources are of growing interests from both a business and an environmental point of view. Both of these tasks require accurate predictions of bike rental behaviors at least on the daily level.

(further motivates & applications?)

In this project, we strive to answer the following question:

- If we have the past history of bike rental counts as well as records of environmental and seasonal conditions, how and how well could we predict the bike rental counts in the future?
- In particular, how and how well could we predict for the next whole year, and what about for the next few days?

Such questions are approached by predictive modeling of daily bike rental counts from a 2011-2012 Bike Sharing dataset [1]. The modeling approach is based on Multiple Linear Regression (MLS), and the daily bike rental counts are predicted using the environmental variables (e.g., weather conditions) and seasonal variables (e.g., holiday schedules) as predictors.

## 2 Background

The aim of this project is to achieve the best model(s) that can be obtained from past data for the use of predictions for the future, preferably predictions one year ahead. To validate the prediction power of models under this setting, the basic goal of this project is to train all models using only the 2011 data, and then test them on the 2012 data.

The response variable to be predicted is the **daily** bike rental count. In the dataset being studied [1], the following 3 types of bike rental counts are recorded:

1. the count of bike rentals by **casual** users
2. the count of bike rentals by **registered** users
3. the **total** count, which is the sum of casual count and registered count.

Two main types of predictors are included in the dataset, the environmental ones and the seasonal ones:

1. environmental variables

(Table 1: A sample of the data - variable names, meanings, units, sample values)

2. seasonal variables

(Table 2: A sample of the data - variable names, meanings, units, sample values)

# 3 Modeling & Analysis

## 3.1 Pre-processing

### 3.1.1 Type Conversion

To be noticed, the value of categorical variables indicates type labels and has very limited physical meaning in the magnitude of those values, which thus cannot be used in the same way as the numeric variables in MLS models. The categorical variables therefore needs to be recognized before the actual modeling process and to be carefully handled.

The below variables are interpreted as Boolean variables and are transformed into `logical`-type variables in `R`:

- `holiday` (holiday or not)
- `workingday` (working day or not)

The below variables are interpreted as categorical variables and are transformed into `factor`-type variables in `R`:

- `season` (season, from 1 to 4)
- `yr` (year, from 0 to 1)
- `mnth` (month, from 1 to 12)
- `weekday` (weekday, from 0 to 6)
- `weathersit` (weather type, from 1 to 4)

### 3.1.2 Value Conversion

The recorded values of `temp` (measured temperature), `atemp` (feeling temperature), `hum` (measured humidity) and `windspeed` (measured wind speed) in the data set being studied here are the normalized ones; all recorded values are the ones that have been divided by the maximum of measured values [1]. For example, the recorded values of `temp` (measured temperature) are obtained by dividing the original measured values by 41 (max) and are thus all less than or equal to 1.

In this project, these normalized records are scaled back to their original values for the sake of easier interpretations. For example, the recorded values of `temp` (measured temperature) are multiplied by 41 (max) in the pre-processing process, which recovers the original scale of temperatures in Celsius.

## 3.2 Variable Selection

### 3.2.1 Response Transformation

Notably, the behaviors of rental counts from different user types are considerably different.

1. **Patterns with weekdays** (see Figure 1,2): Over the time span of a week, the casual count usually reaches its minimum in the middle of a week (grey dots mostly) and its maximum on weekends (green dots mostly), while the registered count does the opposite.
2. **Patterns with temperatures** (see Figure 3): The casual count seems more linear in both the feeling and measured temperatures (`atemp` and `temp`), while the registered count seems to be (at least) quadratic.

We therefore expect that the registered counts and casual counts will follow different distributions and should thus be predicted by separate models. Furthermore, for the casual count, avoiding unnecessary higher other terms has the benefit of more stable computations and model structures. The prediction of total counts will then be obtained by adding the predicted registered counts and predicted casual counts together.

### 3.2.2 Predictor Selection

Given the predictive nature of modeling in the current problem setting, the predicted response is of greater interests than the actual value of the parameter estimates, as opposed to that in an inference task. This, to some degree, relaxes the constraint forbidding colinearity in the predictors, since colinearity will only lead to instability in the parameter estimates but not in the predictions; however, we should still seek to minimize colinearity at least in our beginning model, which would lead to clearer model structures as well as better interpretability of model statistics at the early stage of modeling, which could provide us clearer directions in the improvement process that follows.

With the above considerations in mind, the predictors in the beginning model are selected following the 2-step approach below:

1. The scatter plot matrix for the whole set of variables are plotted for the 2011 training dataset, and all predictors that seem to be significant, i.e., predictors with which the response variable (daily rental count, `cnt`) exhibits a notable visual pattern, are selected.
2. From the selected predictors above, all the highly correlated predictors are removed. Within a group of correlated predictors, only the one that has the largest correlation coefficient with the response variable as well as having the strongest causal relation with the response (in the intuitive sense) will be kept.

It is important that the investigation is done for all predictors for the sake of minimal loss of information. Note that in practice, the whole set of predictors is divided into two groups, environmental and seasonal, and plotted separately, for better readability of the large scatter plot matrices. The separation is justified by the fact that most environmental variables, such as weathers, are expected to be independent of the seasonal variables, such as weekdays and holiday schedules.

At last, the above process leaves us with a small subset of the very core predictors for our beginning model: `weathersit`, `atemp` and `weekday`.

## 3.3 Initial Modeling

In the model building and selection process, we start from the simplest models, which have the minimal number of predictors all in the additive form, as the beginning models.

### Beginning Models

1. For **total count**:

$$\text{cnt} \sim \text{wkngday} + \text{weathersit} + \text{atemp} + \text{atemp}^2$$

2. For **registered count**:

$$\text{reg} \sim \text{wkngday} + \text{weathersit} + \text{atemp} + \text{atemp}^2$$

3. For **casual count**:

$$\text{cas} \sim \text{wkngday} + \text{weathersit} + \text{atemp}$$

(Table 3, 4, 5: eval tables for total, registered and casual)

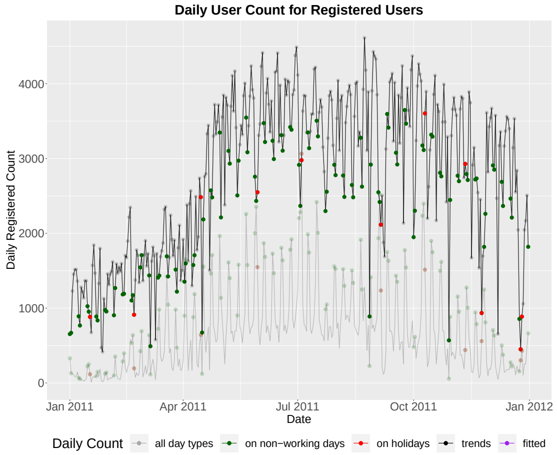
| Model | rmse | n-rmse | % Error | cv-rmse |
|---|---|---|---|---|
| 2011 cas | 309.97 | 0.56 | 72.87 | 314.86 |
| 2011 reg | 584.55 | 0.52 | 25.37 | 594.36 |
| 2011 tot | 722.15 | 0.52 | 25.44 | 734.18 |

Table 1: Diagnostics for Model 1

In this process, the model statistics such as p-values are not relied on as much, because the colinearity issues worsen, which might weaken the significance of

### Final Models

## 3.4 Diagonostic Analysis



interpretation for the final model as well as residual diagnostics

## 3.5 Validation and Problemshooting

## 3.6 Prediction of the Yearly Growth Ratio

The modeling is based on the assumption that the growth trend will remain the same in the future years as that in the year of 2011. Note that this is NOT saying that the user base is supposed to remain unchanged throughout the entire year; the fact that the same scaling factor works at all points in the entire year is due to the fact that the MLS model in the later part of the year, e.g., in fall and winter, are already trained to compensate for rental count growth due to user growth using the environmental and seasonal variables. As such, an estimation of the growth ratio, which we call $\hat{g}$ is very important and worthwhile. Its true value, by construction is well estimated by the ratio $\frac{\mathbb{E}_{2012}(C)}{\mathbb{E}_{2011}(C)} \approx (0.608)^{-1} = 1.64$, which is the average counts in 2011 divided by the average counts in 2012. Note that this means our user base grews by around 64% between 2011-2012, which is a non-trivial amount. If unaccounted for, that magnitude of growth would (and does in fact) lead to chronic model underfitting. However, since we are not **allowed** to peek at the 2012 data, we must resort to other creative techniques in estimating $\hat{g}$. In this report, we propose two techniques for estimating $\hat{g}$: the environmental loss function technique (more involved) and the window technique (more simple).

### 3.6.1 Environmental Loss Function

The primary issue of relying solely on one years' worth of data to estimate the growth within that year is that there is great difficulty in factoring out the role environmental factors play in determining the difference in observed counts between any two given days. As such, if we are somehow able to find two days whose weather/environmental conditions are "similar" or "equivalent", the ratio of the counts between the later day and the earlier one is a reasonable "data point" useful for obtaining a sense of the growth. This follows because we asked those two days to be "environmentally equivalent" (we will formalize this notion soon), so *a priori*, if we were somehow able to marginalize out the environmental factors, then any observed difference in the count `cnt` **must** be attributed to user base growth! This motivates our technique, aptly named the "environmental loss function" technique. That being said, let us slowly develop our loss function idea.

Suppose we have two days (observations) $d$ and $d'$. Consider the reduced dataset where the observations are solely the vectors of our *continous environmental variables*, in our case, we have three: `atemp`, `hum`, and `windspeed`. Furthermore, suppose that we **standardize** these variables. This is critical since we don't want our loss functions' units to be arbitrarily inflated/deflated. This is one of the primary reasons we opt for a continous variable. Again, we standardize the variables so the loss function is unit-agnostic. That being said, this means we may write $d = (x_1, x_2, x_3)$ and $d' = (x'_1, x'_2, x'_3)$ where $x_i$ represents the standardized value of predictor $i$.

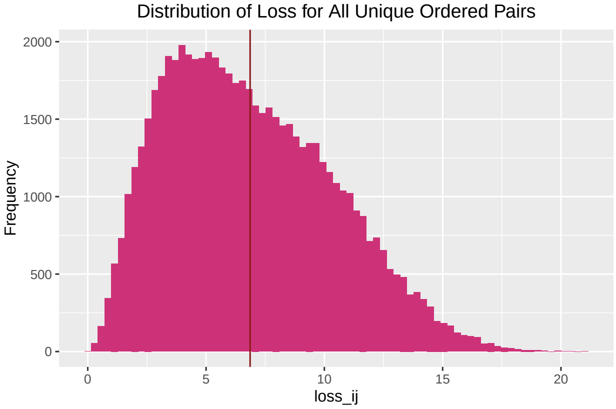
Also, let's use `holiday` to focus only on non-holidays and removing holiday days from the dataset. We can adjust for environmental factors but comparing two "environmentally equivalent" days without accounting for `holiday` may very well throw off our estimates of $\hat{g}$. Losing the 11 days which are holidays prove to be trivial, since $\binom{365}{2} \approx \binom{354}{2}$. Note that we omit `weathersit` and `workingday` (albeit not environmental, is still important) since they are categorical variables. While loss functions can include categorical variables, we will opt for the simplicity, familiarity, and stability of continous variables. In any case, we can justify their omission anyway since `weathersit` is correlated to our chosen environmental variables, so the information loss is not catastrophic. Additionally, since we are using `cnt` and not the subdivisions of `cas` and `reg`, the explanatory power of `workingday` is "canceled" out in the `cnt` variable in a sense since we observe inverse behavior of `cas` and `reg` with respect to `workingday`. Now, we may put the discussion of chosen predictors aside.

Furthermore, let $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ be a weight vector, more heavily weighing variables deemed more important in determining "environmental similarity". We deem this to be `atemp`. In practice, we end up weighing `atemp` by $\alpha_1 = \frac{3}{5}$ and the other two predictors, `hum` and `windspeed`, by $\alpha_2 = \alpha_3 = \frac{1}{5}$. Note that this is arbitrary, and subject to scrutiny, but for the sake of simplicity, we may all agree temperature is the major determinant in constituting a notion of "environmental equivalence". Finally, let $\delta_i$ denote the weighted difference between in the values of predictor $i$ between day $d$ and $d'$. So, we write $\delta_i = \alpha_i(x_i - x'_i)$. Observing the differences among all our predictors, we obtain the difference vector $\vec{\delta} = (\delta_i)_{i=1}^{3}$. Finally, a very natural notion of "loss function" can be defined, namely, the magnitude of the difference vector! We restate this below. Consider the loss function $\mathcal{L} : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$, which is simply the magnitude of the weighted differences of our normalized predictor variables between our days. We write the *environmental equivalence loss* of $d'$ with respect to $d$ as follows:

$$\mathcal{L}_d(d') = |\vec{\delta}| \text{ where } \delta_i = \alpha_i(x_i - x'_i)$$

Now, suppose we enumerate all possible unique pairs of days, which we denote $\Pi = \{\pi_{ij} = (d_i, d_j) \mid i < j\}$. Note that we use the lower-triangle scheme, where we choose $i < j$ so that $i - j < 0$. This will prove useful later on since we want day $i$ to preceed day $j$. We call this the "lower-triangular" scheme because this corresponds to the unique pairs of indices found in a "lower-triangular matrix". In any case, we compute $\mathcal{L}(\pi_{ij})$ for every pair and call it $l_{ij}$. Then, we obtain the set (which we call `df_loss`) $\mathcal{L}(\Pi) = \{l_{ij} = \mathcal{L}(\pi_{ij}) \mid i < j\}$. This is an exhaustive and computationally expensive procedure, since we compute the loss for $\binom{354}{2} \approx 62,000$ pairs of days! So, in practice we compute this once and store the data in `df_loss.csv`.

Now, here is the magic! We are not interested in taking pairs $\pi_{ij}$ which are not very environmentally similar. After all, our quest is to estimate growth between days where we in theory expect a similar amount of users through the marginalization of environmental factors. So, there comes a question, which pairs do we discount as days too disparate to be considered environmentally equivalent? Well, our good foresight in normalizing the continuous variables and weighing them properly (to the best of our abilities) means that we can immediately consult distribution of $l_{ij}$, which we see below:



A quick note: it should not be surprising that we obtain a distribution that resembles a $\chi^2$ distribution. After all, since we assume our predictors are (roughly) normal, their differences will be too; since our loss is essentially the norm of a multivariate normal, this explains our distribution shape! In any case, since we have done a good job by normalizing our variables (making our loss unit-less), a valid strategy to **discount all pairs above a pre-specified upper bound**, which we will call $B$. An appopriate cutoff anywhere below the mean of our distribution $\mu \approx 6.86$, shown as the red line above. For example, we can take $B = \mu - 1\sigma \approx 3.35$ to be a reasonable cutoff for our upper loss bound $B$.

So, to summarize, we want to vet our exhaustive list of all possible day pairs $\Pi$ and throw out bad pairs after doing some "quality control" to obtain a filtered set of pairs $\tilde{\Pi}$. Let us take a brief moment to recap everything. We are interested in estimating $\hat{g}$, and one way we are interested in doing so is observing the growth in pairs of days where the environmental effects are marginalized. A strategy for doing so is to consider "reasonably good" pairs of days, which we explicitly quantify through filtering $\Pi$ through an upper bound on the loss $B$, which we may holistically choose based on the reasoning above.

So, we obtain our filtered/selected set of day pairs $\tilde{\Pi}$, by cutting off the maximum loss at the selected upper bound $B$. That way, we can write our selected pair set as a function that filters $\Pi$ with respect to the chosen loss upper bound $B$:

$$f_{loss}(\Pi, B) = \tilde{\Pi}_B = \{\pi_{ij} \mid l_{ij} = \mathcal{L}(\pi_{ij}) \leq B\}$$

Next, we also filter for what we call an "index difference bound". This is our second round of "quality control". Namely, the idea is simple: we don't want to pick days too close to each other. We know days close to each other will have similar environmental conditions, but as to avoid pesky autocorrelation and vaccuous/misleading information, we impose a bound on how close the days can be.

Let $\rho := \min(i - j)$. In our lower-triangle matrix analogy (where we imagine indices as positions of matrix entries), $\rho$ denotes which diagonal band we are considering. In any case, by imposing a lower bound on $i - j$, to pass the index difference filter, every pair must be atleast $\rho$ days apart, or in other words, be at least $\rho$ diagonal bands away from the corner or towards the main diagonal. In practice, we find low values of $\rho$ to produce the most impactful filters, this makes sense, owing to the autocorrelation observed in days close to each other mentioned above. This makes the index difference bound much less critical than the loss bound, as it is much more impactful for low values of $\rho$. So, to summarize:

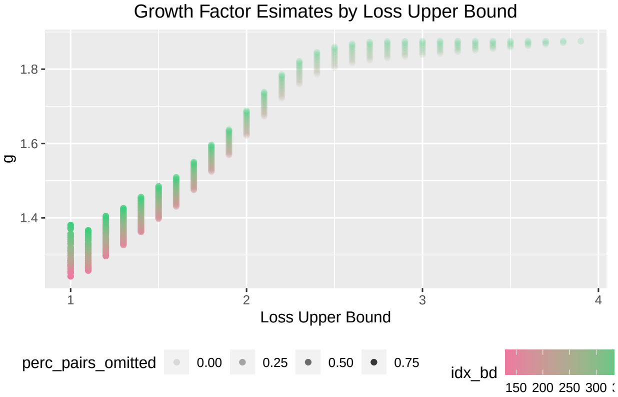
$$f_{idx\_diff}(\Pi, \rho) = \tilde{\Pi}_\rho = \{\pi_{ij} \mid i - j \geq \rho\}$$

Finally... we can compute the estimate derived from the filtered pair set $\tilde{\Pi}$ as follows. Let $c_i$ denote the count of bike users in day $i$, and $c_j$ those in day $j$. **Since we have, in theory, marginalized out the environmental factors (and holidays), we may obtain a reasonable estimate for g by taking the observed growth observed between day i and day j, then taking the average over all the pairs in our filtered pair set.** Since by construction, we used lower-triangular indices, this means day $j$ is always after day $i$. As such, we can imagine each pair, which has marginalized environmental factors, to contain true information about user-base growth since day $j$ is "environmentally equivalent" to day $i$ and is in the future of day $i$! As such, we could in theory attribute this growth to none other than the user base growth itself. To be more explicit, consider a filtered set of pairs $\tilde{\Pi}_{B,\rho}$. Since every pair is considered "good", from each pair $\pi_{ij}$ we obtain a preliminary estimate which we denote $\gamma_{ij} := \frac{c_j}{c_i}$. In general, the $\gamma_{ij}$ are slightly unstable, and their variance may be worth studying; this remains outside the scope of the paper. However, since we are aggregating pairs from a set of $n = 62,500$ pairs (in practice we choose less), there is an abundance of data to counterbalance this. So, to summarize, given a filtered set of pairs $\tilde{\Pi}_{B,\rho}$, we obtain $\hat{g}$ as by taking average of the set of estimates $\gamma_{ij}$ from every $\pi_{ij} \in \tilde{\Pi}_{B,\rho}$:

$$\hat{g} := \mathbb{E}[\gamma_{ij}] \text{ where } \gamma_{ij} = \frac{c_j}{c_i} \text{ and } \pi_{ij} \in \tilde{\Pi}_{B,\rho}$$

**Estimator Analysis in Bound Paramaters Space.** That being said, with our methodology explained, how does one estimate $\hat{g}$? This is our goal in the end! Well, as suggested by the notation above, the parameter estimate $\hat{g}$ is dependent on our filtering thresholds $B$ and $\rho$. Of course, any choice of $B_0$ and $\rho_0$ would be arbitrary, so our next course of action is to observe and analyze our estimator over the bound parameters space. To be concrete, we say fix some ranges $[B_0, B_1]$ and $[\rho_0, \rho_1]$, then observe the behaviour of the resulting parameter estimate $\hat{g}$. To do so, we use the `df_loss` previously mentioned in conjuction with the `get_df_param` function which takes these ranges as an input and outputs `df_param`. A glimpse into what our table looks like is shown below. Note that $n$ counts the number of pairs that "survive" the filtrations given the thresholds. In principle, we want our thresholds to pick exclusive pairs of days which are *unusually* environmentally similar (low $B$) but far apart enough (high $\rho$), with emphasis on the former. That being said, we now plot the results from this dataframe and obtain the exciting results below!

```
##   idx_bd loss_bd        g    n
## 1    130     1.0 1.242667  521
## 2    138     1.0 1.252098  506
## 3    138     1.1 1.258080 3198
## 4    146     1.0 1.256074  497
## 5    146     1.1 1.260398 3149
## 6    146     1.2 1.297205 8044
```



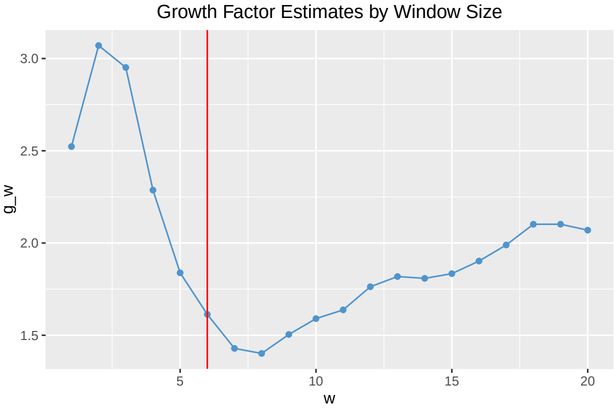
Again, as mentioned previously, the index difference bound is more sensitive when it comes to lower values. Greener points suggest that the index difference bound is high, so our days are quite far apart. Before we delve into the behaviour of the loss upper bound $B$, note that the opacity corresponds to the "percentage of pairs omitted". Recall that we want an exclusive set of pairs in which few pairs survive our strict filtrations or quality control. This is what the plot is suggesting, the more "faded" estimates of $\hat{g}$ are less well-founded, in fact completely unfounded as $p \to 1$ since we would indiscriminately be considering all pairs! As such, we conclude that the optimal range for $B$ probably lies between $[1, 2]$, as these points have more "clarity" in their predictions due to their exclusively low loss. Otherwise, we also notice an interesting pattern with the index difference bound which makes sense, allowing days closer to each other will suggest "less growth". That being said, this means that in some sense, a potential optimal range for $\hat{g}$ could be that around the values where its variance starts to stabilize with respect to $\rho$. Visually, this is the range where the "length" of the streaks starts to hold still. For values before the critical $B \in [1, 2]$ range, this corresponds to the values around $\hat{g} \in [1.4, 1.6]$. That being said, this analysis is holistic, and not 100% conclusive. However, seeing this continous picture of estimating $g$ and narrowing down its true value to be close to $\hat{g} \in [1.4, 1.6]$ is not bad!

### 3.6.2 Method II: The Window Technique

Jeez, that stuff was too technical. Thankfully, the window technique is much simpler. Remember our goal is to estiamte $\hat{g}$, which is the growth rate from 2011 to 2012. As such, the most elementary method to use based on this principle is to take the ratio of the counts in the first day of 2011 and those in the last day. This is justified because everyone would agree, there is some notion that Dec 31, 2011 $\approx$ Jan 1, 2012. However, the issue with this is that our estimate is based on one pair of points ($n = 1$).

As such, to preserve the spirit of the principle, but generalize as to instill more certainty and stability in our estiamte, we describe the "window technique". Firstly, we will say that the method mentioned above is the window technique, for a window size of $w = 1$. The principle of the window technique is the same as for the method described above, the days at the end of the year of 2011 will be similar to those in the beginning of 2012. So, the best we can do is take an equivalent "window" or sample of days from both the beginning and end of the year of size $w$. This way, we use more data and hence have more stability. Then, using those selected $w$ days, we compute the average `cnt` in the beginning (first $w$ days) and end (last $w$ days), then take the ratio of those averages to estimate $\hat{g}$. We call this estimate $\hat{g}_w$. In our first example, we were describing $\hat{g}_1$.

In any case, just like before, $\hat{g}$ is an estimator based on a parameter, in this case $w$. We must note that as $w$ increases, our assmuptions about "day similarity" between the end of 2011 and start of 2012 slowly starts to weaken. As such, we cannot go crazy and choose large $w$. Below, we exhaustively compute $\hat{g}_w$ for $w = 1, 2, \ldots, 20$, and plot our estimates with respect to $w$.



There are a few things to consider. As expected, the values $w \in [1, 4]$ exhibit highly unstable behaviour in estimating $\hat{g}$. However, it starts to stabilize around $w = 6$, highlighted in red. As it turns out, $w = 6$, holistically speaking is our best $w$ for a few reasons. Firstly, going beyond $w = 6$ would mean including Christmas, which is a holiday and would thus throw off our estiamtes. Furthermore, the values of $\hat{g}$ start to stablize reasonably well around that value, indicating the marginal benefit of increasing $w$ is almost fully realized at this point; remember, our assumptions about day similarity break quickly as $w$ grows. Note that in the plot below, we compute $\hat{g}$ after omitting Jan 3rd, due to its high environmental loss value when paired with the other values. Lastly, one reason we consider $w = 6$ to be a good candidate is the fact that the estimate $\hat{g}$ starts to steadily increase again, indicating suspicious behaviour which could be attributed to the fact that as the window increases, we include days whose temperatures are changing in potentially different directions (generally both warming).

That being said, there is one curious piece of information corroborated by the environmental loss function technique. Namely, both graphs seem to suggest discounting the values $\hat{g} \gg 1.6$. In the second graph, this is suggested by the vacously increasing value of $\hat{g}$ with respect to $w$ starting around $w = 10$, indicating that the point in which there is "good" information about $\hat{g}$ probably lies before that value of $w = 10$, corroborated also by the simple principle that smaller $w$ is generally better in terms of upholding assumptions. So, with all this information considered, we settle for $\hat{g} := \hat{g}_6 \approx 1.613$, which is the most compatible value from our holistic analysis of both of our estimation methods.

## 3.7 Refined Model

### 3.7.1 Prediction with the Yearly Growth Ratio

### 3.7.2 Prediction without the Yearly Growth Ratio

# 4 Prediction

## 4.1 Unadjusted Model

## 4.2 Refined Model

# 5 Discussion

Models for both long-term and short-term predictions are included.

To be noticed, at least one more year's data is needed for a final validation of the refined model, which is not available for the moment. This is to be left for the future work.

Time series

# 6 Appendix

## 6.1 Preprocessing

### 6.1.1 Type Conversion

(codes here)

### 6.1.2 Value Conversion

(codes here)

## 6.2 Variable Selection

### 6.2.1 Predictors Selection

### 6.2.2 Predictors Selection

### 6.2.3 Response Transformation

## 6.3 Initial Modeling

### 6.3.1 Beginning Model

### 6.3.2 Final Model

## 6.4 Diagonostic Analysis

## 6.5 Validation and Problemshooting

## 6.6 Refined Model

### 6.6.1 Prediction of the Yearly Growth Ratio

### 6.6.2 Prediction without the Yearly Growth Ratio