ORIE 4999: MAIL Final Deliverable Spring 2024

Team Members: Logan Abramowitz (la437), Mary Kolbas (mck86), Raina Mandayam (rcm328)

Table of Contents

Introduction and Motivation

The MAIL team of ORIE 4999 focuses on using data science and simulation modeling to help reduce service times at the Robert Purcell Community Center (RPCC) Mailroom, the student mail center for residents on North Campus. Our team serves to build on previous semesters' research and discover empirical evidence of theoretical benefits through simulations to ensure promising adjustments to mailroom operations.

In particular, our team focused on exploring proposed changes to the package sorting process that could decrease the time required for employees to search and find a student's package when they arrive at the center, in turn also reducing the queue and wait times for students. We developed and simulated Python simulations for the existing sorting method by last name, as well as the proposed Hash function sorting method, to quantify the benefits of making this change. We engaged with Brandi Smith-Berger, Director of Community Center Operations, to incorporate real parcel report data, ensure our models were reasonable, and discover new ways to engage with the everchanging mailroom operations.

Overall, through this research experience, the MAIL team was able to gain a deeper understanding of the operations of the RPCC mail center. The team learned a great deal about simulation modeling, data science, teamwork, and communication with an external client in a continuously developing operations space.

Prior Work and Topics Learned

Throughout the semester, we referred to the research completed by the Fall 2023 Mail Service Team. The team collected over a thousand data points through observations of mailroom operations and developed a comprehensive queuing model, transitioning from an M/M/N queueing model to a more sophisticated one. They identified peak times for mail center traffic and highlighted inefficiencies in package retrieval processes. To address these issues, they proposed a new storage system using hash functions for package organization, replacing the traditional alphabetical sorting method. This approach aimed to reduce service times and improve overall efficiency. Additionally, the team proposed using the StarRez portal for better package tracking and storage management. Their work provided valuable insights and a foundation for our enhancements in the mailroom's operations.

We learned many valuable skills through this project such as how to effectively use a large spreadsheet of data to model relevant information. This included package arrival rate, student pickup rate and more. We also learned and expanded our skills of python, specifically numpy and pandas where we were able to analyze the simulated data into meaningful visualizations. Lastly, we learned how to create a simulation that models a real-world situation from scratch using python. These skills were acquired through a variety of different sources such as the Fall 2023 Mail team's research and many websites.

Relevant Methodologies and Technologies

StarRez is a comprehensive portal used by Cornell for managing various aspects of housing and mail services. It is integral to the operations at the mail center, specifically in handling package tracking and processing.The StarRez portal is used to check in packages with a student's netID and scan packages out of the mailroom. It helps in verifying student information, especially when package labels are unclear. Then, StarRez automatically generates a unique identifier (UID) for each processed package. This UID is crucial for the proposed new system of package storage and retrieval. The portal offers a service called Data Subscription, allowing the system to assign custom values to packages, add searchable fields, and trigger specific actions.

Simulation

Parameters and Distributions

The simulations both intake the following parameters:

`lastname_dist`: The distribution of starting letters of last names, used to generate student objects.

The `lastname_dist` was taken from Data Mining DNA [1], which compiled the information using the 2010 US Census. The distribution can also be found here: Distribution of First Letter of Last Names

`num_bins`: The static number of bins/"cubbies" in the mailroom. (There may be several physical bins to hold large amounts of packages, but they are all within one "cubby space.")

`arrival_counts_edited`: The distribution of package arrivals based on day, used to determine how many packages arrive on a given day.

`pickup_dist`: The distribution of package pickup. Used to determine the likelihood that the package is picked up on a given day, given how long the package has been in the mail room.

> The `arrival_counts_edited` and `pickup_dist` distributions were taken by aggregating information from the `Copy of Parcel_Report_June_23_22_10_31_23_No_Student_Info (1).xlsx` dataset. See the appendix for details on data cleaning/filtering.

`days`: number of days to run the simulation

Simulation Structure

The simulation revolves around the addition and removal of Package objects from the mailroom. They are added to the simulation at the start of the day based on `arrival_counts_edited`, are assigned to a Student object, and sorted into the appropriate Bin object based on the sorting method (see below). Based on how many objects are in the Bin, Packages are assigned a `location_in_bin` value.

To simulate the day, every package uses the number of days it has been in the mailroom (`days_on_shelf`) and the `pickup_dist` to determine its likelihood of being picked up during this particular day.

If the package is processed, the simulation determines how much time it takes to process based on its Bin (`rackTime`, time it takes to walk to the bin), it's `location_in_bin` (`findTime`, how long it takes to find the package within the bin), and total processing time (`totalTime`, `rackTime`+`findTime`).

After all packages that were selected to be picked up are processed, attributes of the Package objects, such as how many days they have been in the mailroom, are updated. If a Package has been in the mailroom for over 7 days, it is removed from the simulation.

The simulation outputs two data frames at the end of the simulation:

> `df` records a row for every package picked up, which includes processing time and what day the package was processed.

`df_total_packages` records a row for every day of the simulation, including data on how many packages remain in the room at the end of the day and how many were removed for being over 7 days old.

Last Name Sorting vs. Hash Function Sorting

The only difference between our Last Name Sorting simulation and Hash Function Sorting simulation occurs when packages are added to the simulation at the start of each day.

The Last Name Sorting method places a package based on the assigned student's last name. In our tests we used 26 bins, meaning each letter had its own bin. Due to the non-uniform last name distribution, we would expect certain bins to have more packages than others. Bins in this simulation do not have a maximum capacity.

The Hash Function Sorting method places a package based on its unique ID, which is incremented based on how many packages have been added to the simulation. The Hash Function is a simple modulo operation `counter_packages%num_bins` that ideally spreads out the packages evenly among the bins. Bins in this simulation have a maximum capacity of 300, and if a bin is full, the Hash Function uses linear probing to place the package in the next available bin.

As a result, we expect certain measures we compare between the two simulations to be relatively similar (only differing due to noise such as randomness of package pickup), while other measures may differ significantly due to the sorting method. Identifying the impact of changing the sorting method is important in accomplishing our goal for this project.

Results

We chose to run both simulations with 26 bins and for 90 days.

Our results can be found in Table 1:

	Hash Model	Last Name	Hash - LastName
Time to walk to rack (units of time)	13.534248	11.388086	2.146162
Time to find package within bin (units of time)	76.130943	117.975499	-41.844556
Total Time (rack + find) (units of time)	89.665191	129.363585	-39.698394
Days for student to pickup package	0.76082	0.755013	0.005807
Average number of packages in the system at the end of the day	1797.9	1792.66	5.24
Total Number of packages returned	15551	15533	18
Average number of packages returned	172.78	172.58	0.2

Table 1. Output Measures of Hash Model and Last Name Simulations,
Hash v Last Name Results

We see all major differences in the measures for processing time, particularly the significant reduction in the time needed to find the package. This matches the hypothesized findings that the Hash Function sorting method would better spread out the packages more evenly among bins, meaning the average time spent sifting through packages would be lower.

We also see that the time to walk to the appropriate bin is slightly longer for the Hash Function. We hypothesize that this is because last names with letters that start near the beginning of the alphabet are more popular, meaning the bins that the service center workers must walk to tend to be closer in the Last Name sorting model.

As expected, measures related to package pickup or the number of packages at the end of the day are very similar, only differing slightly due to the randomness of package pickup.

When comparing the distribution of total package processing time, Figure 1 shows that the Hash Function shows significantly more consistent results, as the package times are clustered more closely around the mean/median. The Hash Function exhibits lower mean and median processing times and is slightly left-skewed, while the Last Name model is more heavily right-skewed.

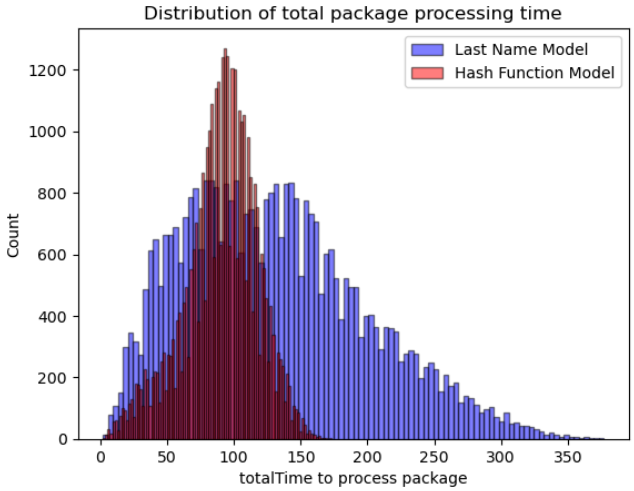

Figure 1. totalTime distribution of all processed packages

We reach an equilibrium of packages in the mailroom and packages expired around day 50 of our simulation, which corresponds to the beginning to middle of September (50 days after August 1). At equilibrium, the mailroom has approximately 2000 packages at the end of each day and sees about 200 packages reach 7 days without being picked up. This 50 day waiting period for the system to reach equilibrium is expected as at the beginning of the semester students often order many more packages than they do the rest of the year. This is often due to students realizing that they do not have something that they need once they have moved in.

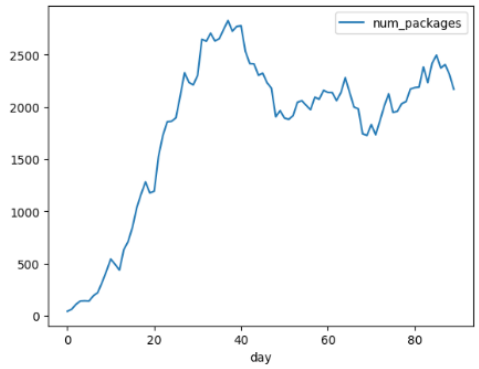

Figure 2. Packages in the mail room at EOD (Last Name Model)

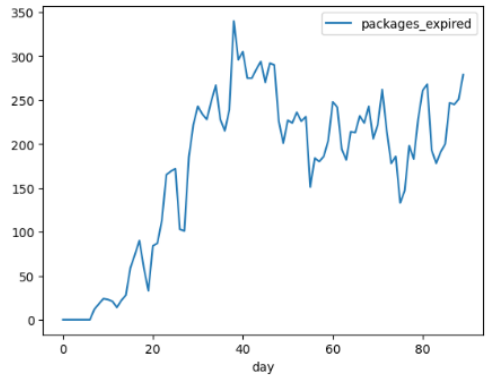

Figure 3. Packages expired at EOD (Last Name Model)

Although both Figure 2 and Figure 3 are based on the Last Name Model, no matter the sorting method these results are similar, only differing in minor noise of randomness in pickup distribution.

A major limitation of this equilibrium analysis is that our simulation starts on day 0 with no existing packages, so the simulation needs several days to acquire a significant amount of packages to be representative of reality. Although over the summer there are significantly fewer students using North Campus mail facilities, lingering packages are common.

We also find in our simulation that the number of expired packages per day is much higher than the actual values of returned packages based on the parcel report provided (Figure 4). This is likely because the August-October time period often has an influx of packages and the mail center tends to be much more lenient given students are arriving and adjusting to campus. Additionally, we hypothesize that students who are being sent packages by relatives are much less likely to pick them up in comparison to the packages they ordered.

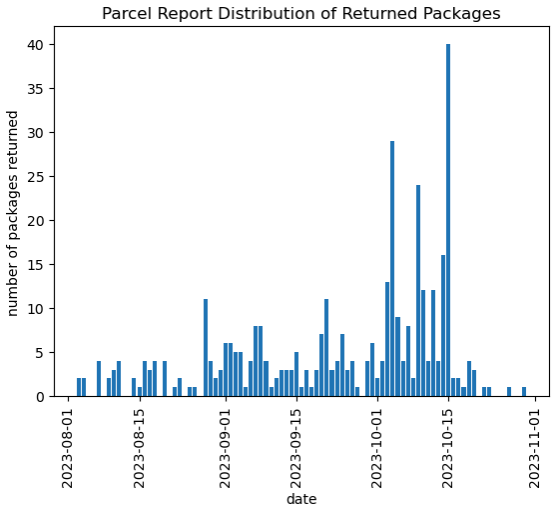

Figure 4. Actual Returned Packages from Fall 2023

Simulation Limitations

Distribution Generalizability: The `arrival_counts_edited` and `pickup_dist` distributions were taken using Fall 2023 Parcel Report data, which means our simulations can only truly be generalized for predicting the impact of these sorting methods on package arrival and pickup behavior present at the beginning of the academic school year. Although this is a limited scope, it was determined to be most relevant to our client, as it is often the most difficult mail influx to deal with.

Distribution Time Limit: Due to the nature of how the distributions are taken directly from the Fall 2023 Parcel Report, the simulation only has enough data to run for up to 91 days. However, if we had a larger dataset continuing past October 2023, this could be easily expanded and implemented. We find that the equilibrium is visible prior to 91 days, so we did not expand our simulation's capacities in this sense.

Starting with no packages: It is not realistic to assume the simulation should start with no packages in the mailroom, but the length of simulation still allowed us to find an equilibrium and make conclusions regarding the impact of high numbers of packages on processing time.

Students with multiple packages: Our simulation does not handle if a student has multiple packages in the mailroom – instead, it assumes each package is unique or they would independently pick up each package based on how long it has been in the mail room. In reality, if a student arrived to pick up package A, they would also pick up package B, no matter how long either had been in the mailroom. Having multiple packages would also likely distort the probability the student comes to the mail center.

No queuing: Our simulation does not calculate student queuing (waiting in line to be serviced) and instead focuses on the processing time behind the counter. We make the assumption that reduced processing times would correlate to lower waiting times. We do not incorporate the probability of students giving up on picking their packages because they do not want to wait. Arguably, but using the `pickup_dist`, some aspect of this "give up" rate is represented.

Next Steps

Hash Function sorting and students with multiple packages: We hypothesize that we can further reduce the processing time for packages if the Hash Function is adjusted to purposefully place packages for the same student in the same bin, regardless of its identification number. However, in order to prove this, the following must also be implemented:

Students with multiple packages: As mentioned in Limitations, in order to make this simulation more realistic, it should be adapted to incorporate how students with multiple

packages would affect the processing times. More data collection would need to be done into how having multiple packages affects the probability of pickup.

Package Size: We incorporated a parameter to the simulation `packagesize_dist` that was not implemented due to project scope and lack of spatial data about the bins and mail room. A further step would be to incorporate package size into the model to be able to understand how differently sized packages could be realistically sorted.

Queuing: Add measures to track how long people spend waiting in line. More data collection would need to be done to understand how likely it is that a person leaves the line after waiting for too long.

Other Avenues Not Pursued

There is another project that has been sidelined which is known as the package lockers. In this project the idea is to distribute packages to students by backfilling lockers and then having the students come and take their packages out of the locker they have been assigned. The student's assignment for a locker would come in the form of an email alerting them that they have a package to pick up, much like the current email that they receive, but it would also include which locker their packages are in. In this system however, students would only have about 3 days to retrieve their packages rather than the normal week like they have now. In theory this would reduce the amount of time a student would have to wait in line to get their package as each student would act as their own server. This project has been put on hold but perhaps with data supporting the potential benefits it can be reinstated.

Github and Google Drive

Github: https://github.com/marykolbas/mailroom_simulations
Google Drive Folder: ORIE 4999 MAIL SP24

Relevant Github Files:

File Name	Description
`mailsimFinal.ipynb`	Final simulations `runLastNameSimulation2` and `runHashSim` Figure 2, 3
`distribution.ipynb`	Creation of `arrival_counts_edited` and `pickup_dist` Figure 4, A1, A2
`mailsim_distributions.ipynb`	Figure 1
`Package Distributions.ipynb`	Exploratory code

`MailSimulation.py`
 	
`mailsim.ipynb`
	Early iterations of simulations

Bibliography

[1] O'Brien, M. (2022, November 30). Most Common First Letters Of Last Names (Statistics). Data Mining DNA. https://www.dataminingdna.com/most-common-first-letters-of-last-names/

Appendix

Data cleaning to create the `arrival_counts_edited` and `pickup_dist` distributions can be found in `distribution.ipynb`. They were created after filtering for only packages in the North Campus Service Center, and the pickup dates were calculated by comparing the package's Receipt Date and Issue Date, specifically for packages that were picked up. Packages that arrived prior to August 2023 were removed, as this is when the 7-day return-to-sender began to be enforced. These distributions are visualized in Figure A1 and Figure A2:

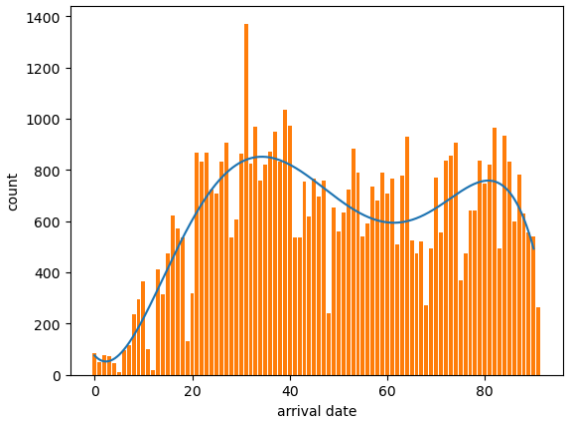

Figure A1. Number of packages that arrive on day d

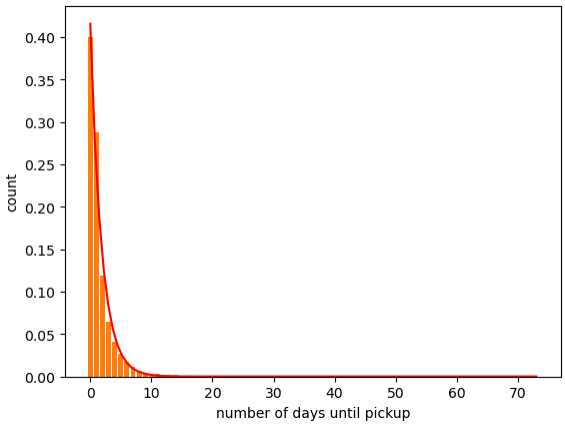

Figure A2. Probability a package is picked up on day