Fund the Vote

Reducing Voter Wait Times using a Predict – then – Optimize Pipeline

Gareth Minson
gminson@andrew.cmu.edu

Raj Shah
rajshah@andrew.cmu.edu

Ashwin Kandath
akandath@andrew.cmu.edu

12/6/24

Attributions

Task Breakdown:

- Data Research and Initial Data Cleaning:
    - Raj looked into identifying Census data, and did initial work on downloading those files, cleaning discrepancies between years, and joining them
    - Gareth looked into the Precinct data, finding elections data in both XLS and PDF format, and joined them together
    - Ashwin worked on finding ShapeFiles for both sets of data, allowing us to merge and combine the full set of data, and worked on the idea for performing weighted averages to determine region-specific demographics

- Hierarchical Prediction Pipeline:
    - Raj developed code to generate interaction features between demographics, then doing LassoCV fits to determine the most integral features.
    - Raj worked on using a hierarchical model with Linear Regression followed by Logistic Regression to predict Voter Turnouts

- Clustering:
    - Gareth and Ashwin worked on utilizing precinct data to best determine the best way to group together precincts in the Pittsburgh area based on demographics.

- Gurobi Analysis:
    - Gareth, Ashwin, and Raj worked on developing the Optimization Model, setting up Decision Variables, and writing out Constraints to best determine the flow between Check-In Stations and Voting Booths

- Results + What If Analysis:
    - Gareth worked on interpreting results and better understanding issues in our model, and developing an alternate model to explore other scenarios.

- Optiguide Analysis:
    - Ashwin worked on using Optiguide to explore several different problems - such as dynamic resource reallocation and sensitivity to prediction data.

Problem Statement and Executive Summary

The goal of our project is to provide a recommendation to Allegheny County on the ***best way*** to minimize wait times and ***improve voting experiences*** for marginalized voters. To do this, our objectives were to:

- Determine a programmatic way to estimate wait times
- Determine the arrival rates for voters in each precinct based on historical data
- Determine which precincts corresponded most to marginalized voters.

To measure these objectives, we decided our final quantifiable metrics would be identifying the number of voters that had to wait in line across all precincts, across all timesteps. We would also identify the number of voters that had to wait in line across "prioritized" precincts, to determine how policy changes affect those that are potentially from marginalized communities.

Using historical data, we were able to predict voter turnout using a Hierarchical Prediction Pipeline - demographic variables were predicted up until 2026 using Linear Regression methods, and then those demographic variables were fed into a Logistic Regression model, which predicted voter turnout with an R^2 score of 0.747. Clustering on demographic variables allowed us to identify precincts of interest, producing 6 distinct precinct archetypes, allowing us to prioritize different marginalized groups.

Our optimization model used this data to allocate a limited number of voting booths and check-in stations across 144 precincts, to minimize the total number of people waiting in five-minute waiting periods throughout election day. Using parameters such as service rates for check-in stations and voting booths, we achieved a solution within 1% tolerance, resulting in a total of 627,005 "wait sessions". On average, voters experienced 6 minutes and 40 seconds waiting to vote. Notably, voters in prioritized precincts experienced a total of 65,024 "wait sessions" - or an average of 2 minutes and 30 seconds per prioritized voter. Notably, all check-in stations were fully allocated.

However, comparative analysis with a neutral model revealed that the primary model only offers nominal improvements for priority precincts, showcasing that more voting booths did not change their wait times. This outcome showed that check-in stations are the predominant bottleneck, and our recommendation to the Allegheny County Elections Division is to focus optimization efforts on alleviating delays at the check-in stage.

Using Optiguide Analysis, we confirmed check-in stations were the primary bottleneck, and explored how doubling them significantly improved queue lengths by nearly 24%. We also explored dynamic resource reallocation with mixed results - overall wait times increased, but service improved in high-flux areas. Finally, the analysis explored how the results from our prediction stage would influence our optimal resource allocation.

Analytical Model and Data

The goal of our project is to provide a recommendation to Allegheny County on the ***best way*** to minimize wait times and ***improve voting experiences*** for marginalized voters. To do this, we decided to formulate the problem as:

$$MinZ = \sum_{i=1}^{I}\sum_{t=1}^{T}\left(W_{i,t}^{c} + W_{i,t}^{v}\right)$$

Where:

- Z is the Optimization problem
- $i \in 1, 2, ...I$ represents each precinct i
- $t \in 1, 2, ...T$ represents each time period t
- $W_{i,t}^{c}$ represents each person in the check-in queue at precinct i and time t
- $W_{i,t}^{v}$ represents each person in the voting queue at precinct i and time t

As such, our optimization problem aims to reduce the amount of people ***waiting*** in queues total across all precincts across all time periods. If less voters have to wait (and are instead immediately serviced), it follows that overall wait times will also be reduced.

The Decision Variables we have are:

- c_i: The Number of Check-In Stations allocated to precinct i
  - $C_{min} \le c_i \le C_{max}, \quad \forall i \in I$
- v_i: The Number of Voting Booths allocated to precinct i
  - $V_{min} \le v_i \le V_{max}, \quad \forall i \in I$

Instead of dealing with a formal budget and allocating costs, we chose to allocate resources - check-in stations and voting booths - to each precinct station, and provide a budget for these resources. Our main constraints that supported the optimization are:

- $W_{i,1}^{c} \ge a_{i,1} - c_i s^c \quad \forall i \in I$
  - Where $a_{i,t}$ is Arrival Rate at precinct i at time t
  - Where s^c is Check-In Service Rate (assumed constant)
  - This constraint defines the initial Queue Length at Check In stations, as a function of the incoming morning arrivals and the amount of check in stations multiplied by service rates.

- $W_{i,t}^{c} \ge W_{i,t-1}^{c} + a_{i,t} - c_i s^c \quad \forall i \in I, \quad \forall t \in \ 2 \le t \le T$
  - This constraint defines the Queue Length at Check In stations, after the initial time step, modifying it to include those leftover in line from the first

time step.

- $W_{i,t}^{v} \geq W_{i,t-1}^{v} + c_i s^c - v_i s^v \quad \forall i \in I, \quad \forall t \in \; 2 \leq t \leq T$
    - This constraint defines the Queue Length at Voting stations, after the initial time step. We add the length of the queue in the last time step, the ones who were serviced in the check in line (from the last time step), and remove the count of those who do vote (given by service rates multiplied by voting booths).

These three dynamics guide the optimization process, controlling how the queues for check-in stations flow into voting booths.

The last major constraint our problem required was for supporting disenfranchised populations. After clustering precincts into different groups, and mapping a Voting Booth Requirement B_d to each cluster d, we were able to have the following: $v_i \geq f(i)$, where $f(x)$ is a function mapping each precinct i to their respective cluster d, then returning the appropriate resource constraint B_d.

The model also had integer and non-negative constraints to ensure all solutions were feasible in the real world.

To develop this model, two sources of information were required - $a_{i,t}$ and d_i - the arrival rate for each precinct at each timestep, and the identifying cluster for each precinct. The following data sources were used to predict these pieces:

- Census Tract Data:
    - **S0601** - Data relating to Demographics, identifies the racial makeup and income levels for Pittsburgh residents from 2014 to 2022.
    - **B18101** - Data relating to Gender and Disabilities for Pittsburgh residents from 2013 to 2022.
    - Both sets of data were cleaned and joined to **Census ShapeFiles**
- Precinct Data:
    - **Elections Data** - Data (CSVs and PDFs) were gathered from here to determine the number of in-person voters on Election nights from 2008 - 2024.

These datasets were merged using geopandas to determine the demographics of the voting eligible population within each precinct. To bolster our analysis, we performed a simple Linear Regression on each demographic variable to estimate how they'd change from 2024 to 2026.

Using those estimates, we were able to perform a more robust Logistic Regression, to determine the Total Turnout Y at each precinct i for each year y.

$$log(Y_{i,y}) = \alpha + \sum_r^R (r_i * \beta_r) + \sum_g^G (g_i * \beta_g) + \sum_n^N (n_i * \beta_n) + \ldots + \sum_r^R \sum_g^G (r_i * g_i * \beta_{(r,g)}) + \ldots$$
$$+ \ldots \sum_r^R \sum_g^G \sum_n^N \sum_p^P (r_i * g_i * n_i * p_i * \beta_{(r,g,n,p)}) + \beta_Y Y_{(i,y-1)} + \beta_m M_y$$

The Regression factored in the racial, gender, income, and disability makeup of a precinct each year, along with interaction terms between every combination of those different groups. It also utilized the turnout from the previous year, and whether or not the current year was a Midterm or not.

Using this framework, we were able to achieve an R^2 of 0.747 - most errors came from heteroskedasticity issues, where the model performed better in precincts with less than 1000 voting people. Using the logistic regression, data for 2026 was able to be generated for the number of voters at a given precinct. This number was then divided into a randomly-noised trimodal distribution, simulating voting patterns throughout the data, and providing us our values for $a_{i,t}$.

To cluster our data, the K-Means algorithm was applied to the latest precinct demographic data, identifying clusters with similar characteristics. Silhouette Scores were used to determine the optimal number of clusters, and each cluster was individually analyzed to determine the characteristics differentiating it from other clusters. This allowed us to “prioritize” certain precincts over others, using their membership within clusters to identify how many resources to allocate to each one.

Our methodology of downloading and cleaning historical data, performing statistical regression on said data, and then using optimization techniques on our predicted outputs, allowed us to provide the county with a recommendation on how they could best spend their resources to minimize voting times.

Post Model Analysis

Our primary version of the model optimizes allocating a limited set of voting booths and check-in stations across the precincts while observing minimum allocation levels and requiring a higher minimum allocation at precincts where we observe a high proportion of typically disenfranchised voters. The model minimizes the sum of voters that wait to check-in or having checked-in, wait to vote over all 144 five minute periods in each precinct over the course of election day.

The parameters of this model are:

- The number of voters that can be checked-in in a five minute period (2.5 or 1 person every two minutes),
- The number of voters that, having been checked in, can vote per period (5 or 1 person every minute),
- The minimum number of check-in stations (2),
- The minimum number of voting stations (3 or 5 in a precinct of concern),
- The maximum number of check-in stations (5),
- The maximum number of voting stations (10),
- The number of check-in stations to be allocated (3,396 or 3 for each precinct),
- The number of polling stations to be allocated (10,188 or 9 for each precinct),
- The predicted arrivals at each precinct in each 5 minute period.

We expect a difference in check-in speed and voting speed for a few reasons. The check-in process requires a staff member finding a voter in their eligible voter list and the voter signing their name in the book. This is two parts, conducted by two people, and we expect inefficiencies or hiccups to occur. The voting process, however, requires filling in bubbles next to the voter’s selection for as many of the races they choose and submitting the form. This process that only requires one person has less chances for events to impede its resolution.

Voting stations are more plentiful as a minimum and at capacity for a similar reason: voting stations do not require additional staffing. Check-in stations availability and provision is effectively a reflection of the staffing capacity of the county election commission.

Importantly, we did not generate these numbers at random, but they were based on anecdotal evidence from interviews we conducted with peers voting at an assortment of precincts and by reviewing the comprehensive polling station manual for Allegheny County.

The model found a solution with a tolerance of 1%. The total 5 minute waiting periods were 627,005. Importantly, this is not the number of people that waited as a person could spend multiple periods waiting either in the check-in line, the voting line, or in combination. Check-in waiting accounted for 53% of waiting periods. On average, the model estimates that a voter spent 6 minutes, 40 seconds waiting in some manner.

Voters in priority clusters waited a total of 65,024 periods or an average expected wait of 2 minutes, 30 seconds per voter. The expected voters in priority clusters account for 27% of all expected voters (469,885) and experienced 10% of all waiting periods.

As for the resource division, all the check-in stations were allocated, but only 89% of voting stations were placed in precincts. With the higher minimum provision of voting stations in precincts of concern, only 32% were given more than 5, their minimum, but 86% of all other precincts were assigned more than 3 voting stations, their minimum.

We interpret these results to indicate that check-in stations are the limiting factor in reducing wait times. There is a higher time penalty at the check-in stage and they have a larger staffing cost.

Ultimately, this model cannot evaluate itself. On its own it does not reveal enough to say whether the priority allocation of voting booths generates improvements for priority precincts. To evaluate that then, we added a complementary model that had the same design and parameters except that all precincts had the same minimum allocation of voting stations (3) for all precincts.

This model found a solution with 1% tolerance. The total expected 5 minute waiting periods was 627,679, marginally more than the first model (674) which did nothing to change the expected waiting time per voter.

Importantly, voters in a priority cluster did not increase their expected waiting time in this model. The differences in the model were in voting station allocation. On average, in the priority model, priority precincts had 6.6 voting stations and all others and 9. In the neutral model, the priority precincts had 6.2 voting stations and all others had 9. The neutral model even allocated more than 5 voting stations to 67 or 44% more priority precincts than the priority model. The neutral model made these adjustments while placing 2% less voting stations in precincts.

The neutral model made marginal changes to check-in station division (less than 5%) and used all check-in stations.

The priority model provides only nominal improvements for priority precincts by allocating more voting stations on average than a neutral model, but offering no change in average expected waiting time.

Considering the results, we recommend that the Allegheny County Elections Division consider optimization strategies that will reduce wait times and consider waiting as an effect of each stage of the voting process. The Elections Division should consider our premise that voting should be fast enough so that everyone can do it, but rather than simply following our allocation guidelines engage with stakeholders in the community about where they believe the lack of investment occurs, take a critical analysis of the recommendations, and implement a solution similar to our methodology. Future work in this policy field should consider travel time, parking availability, weather sensitivity, and more dynamic factors. Our model is limited in how it accounts for mail-in voting which is becoming more popular and perhaps accounting for that will generate more equitable and efficient results.

What-If Analysis:

Since check-in stations are a limiting factor for voting process time because they require additional staff and are slow to move through, as a what-if analysis, we included an additional version of the modeling to estimate a return on investment in check-in stations. This approach uses a two-step optimization and can be found in the appendix. In this model, first we estimate the number of check-in stations required to achieve an average expected waiting time of 5 minutes per voter and second, distribute them with a preference shown to precincts with historically disenfranchised voters.

With a 1% tolerance, the first stage estimated that the total check-in station capacity could be as low as 3,000. This is lower than the initial total capacity of our primary model. The second step, distribution optimization, with 1% tolerance, reduced the expected average wait time to 4 minutes, 56 seconds for all voters and 4 minutes, 20 seconds for voters in priority clusters. Importantly, this model set a higher minimum voting booth allocation and higher minimum check-in station allocation for priority clusters.

Optiguide Analysis:

To further analyze the performance of our model and derive deeper managerial insights from the project, we implemented Optiguide, which is a '*framework that accepts as input queries in plain text, and outputs insights about the underlying optimization outcomes'.* It indeed gave us some interesting insights from our model and did a what-if analysis on various aspects of our methodology. It is important to note that the outcomes of this analysis lie in an error bound of 6% in order to optimally utilize computation resources.

Firstly, we observed that the number of voting booths at each precinct were not as crucial in the optimization of total waiting times as the number of check-in stations were. Thus, we queried to understand how the change in the lower bound for check-in stations (especially in disenfranchised locations would impact the objective function value).

Question 1: Right now, to assist the disenfranchised clusters (1,3,4), we are setting a higher lower bound for the number of voting booths by calculating min_booth_voting. How will the objective function change if we do the same for check-in stations, instead of voting booths?

Response:

> *By setting a higher minimum number of check-in stations for the disenfranchised clusters (1,3,4), the model now accounts for the increased need for resources in these clusters. This means a minimum of 2 check-in stations were allocated to these clusters instead of the original 1 for clusters 1, 3, and 4.*
>
> *However, the objective value (total queue length over all precincts and time periods) remains the same as in the original solution, at 634468. This suggests that the algorithm was already allocating at least the new minimum number of check-in stations to these clusters in order to minimize queue lengths, hence the unchanged objective function value.*
>
> *In conclusion, setting a higher minimum number of check-in stations for certain clusters is a way to guarantee a certain level of service for those clusters, but in this case, it did not change the optimal solution or the resulting objective function value.*

Another important assumption we made while building the model was the allocation of resources happen before the election day. However, we wanted to assess how the wait times change if the allocation was dynamic.

Question 2: How can dynamic reallocation (based on predicted peak times) improve overall system performance?

Response:

> *Absolutely. Comparing the new results with the original, we can see that the objective value has slightly increased from 634468.0 to 635111.0. This indicates that the overall waiting time in queues has gone up a bit with the added constraints for dynamic reallocation of resources.*
>
> *The slight increase in the objective value might be a result of the fact that resources are being moved away from non-peak times to better serve times of peak demand. What this implies is that although the overall waiting time may have increased slightly, the service during peak times might have improved as resources are being more effectively allocated to those high-demand periods.*

Hence, while the objective of minimizing the total queue sizes has not been perfectly optimized, the dynamic resource allocation could lead to improved service during peak periods, which can be seen as a trade-off.

Nevertheless, it's important to consider all factors in a real-world scenario, and whether the benefits during peak times outweigh the increase in overall waiting times would require further analysis and potentially more balanced performance metrics.

Another caveat in our modelling is that we have assumed a trimodal distribution of arrivals since the data for hourly arrival of voters was not available. Thus, we probe more into the arrival process and voter turnout. In the next 2 questions, we explore how would the optimal waiting time behave if there is a change in the turnout.

Question 3: How would increases or decreases in voter turnout in priority precincts impact the optimal allocation strategy?

Response:

Yes, the changes to the voter turnout rates in individual precincts significantly impacted the allocation strategy of checking stations and voting booths.

After adjusting the arrival rates based on the new turnout rates, the optimization model was solved again. The new optimal allocation resulted in an objective value of 151021.0, which represents a metric for the total queue lengths in the system. This is a decrease when compared to the original objective value of 634468.0 prior to changes in turnout rates.

This indicates that with the decrease in voter turnout in certain precincts, our model was able to more efficiently distribute resources. This leads to shorter total queue lengths, ultimately resulting in improved voter service. The specific changes to the allocation of resources would depend on the details of the changes in turnout rates across individual precincts.

In conclusion, the allocation of resources is dynamic and should adapt to voter behaviors for optimal performance. Small changes in factors such as voter turnout rates can greatly influence the optimal allocation strategy. Therefore, it is crucial to regularly update these parameters and rerun the optimization model to ensure efficient resource distribution in the voting system.

Question 4: If turnout predictions are off by 10%, how do total wait times and resource allocations shift?

Response:

> *After increasing the turnout predictions by 10%, the total wait times significantly increased from 634468 to 1145216. This indicates that if the turnout predictions are off by 10%, the queues would be significantly longer, leading to increased wait times for voters. However, please note that the specific changes in the resource allocations (e.g., number of voting booths and check-in stations per precinct) would require further analysis. Overall, accurate turnout predictions are crucial for efficient resource allocation in this context.*

We have also assumed that the number of check-in stations and voting booths available are a function of the number of precincts (by multiplying a constant factor). We now explore how the optimal allocation would change if we had more resources.

Question 5: If additional resources were available, where should they be allocated to achieve the greatest reduction in wait times? Consider doubling the number of check-in stations.

Response:

> *By doubling the number of check-in stations, we've seen the total wait time significantly reduced. Originally, with the former setup, we had an objective value of 634468 (representing the total time that all voters spend in queues in all precincts during all time periods) while with our new setup, the objective value reduced to 485462, representing a 23.5% decrease in total wait time. This suggests that allocating additional resources to check-in stations can indeed effectively decrease the overall wait times in precincts.*

Finally, we explore where our model could do better, and any scope of improvement and further study.

Question 6: In our implementation, do you think we have missed out on something important that can possibly alter the outcome of the model? If yes, give us concrete actionables to improve the model.

Response

Sure, here is a human-readable interpretation of the result:

Based on the optimization model we implemented, the optimal solution leads to an objective value of 634468.0. The objective value represents the total waiting time across all the precincts, i.e., it is the sum of waiting time in both the check-in queue and voting queue for all precincts across all time periods. So, in our current approach, we managed to minimize this total waiting time across all precincts to 634468.0 units of time.

The changes we implemented to optimize our model are configured in the parameters regarding the maximum number of voting booths and check-in stations. By exploring different configurations for these parameters, we found the configuration that yields the least amount of total waiting time.

It's important to compare these results with the results obtained from the model before optimization. The improvement in total waiting time signifies that our model is working as intended and we have been successful in our endeavor to minimize waiting time.

Further optimization could be achieved by considering a more dynamic approach to the arrival rates, considering the distribution of voters among precincts, and allowing for variability in service time. However, with the current model and assumptions, the optimal total waiting time we've achieved is 634468.0 units.

Based on the conversion between the agents:

```
# The model currently does not consider the change in arrival rates over time.

# Arrivals rates are taken from the data, but does not take into account any variability or underlying trends in the data.

# This could be improved by modeling the arrival rates as a stochastic process rather than static values.
```

```
# Further, model also does not consider the distribution of voters among precincts.

# It assumes that voters are evenly distributed which may not be the case in real situations.

# A data-driven approach could be used to estimate the distribution of voters.


# Lastly, the service time is assumed to be constant which may not necessarily be the case.

# Depending on the complexity of the voting process or other factors such as time of day,

# the service time could vary.

# This could be modeled by a separate variable or function in the model.
```

From the analysis using Optiguide, we gathered the following insights.

1. **Impact of Higher Minimum Allocation in Priority Clusters**
   Increasing the minimum number of check-in stations for disenfranchised clusters (1, 3, 4) did not lead to a significant change in the overall queue lengths or system performance. The objective function value remained unchanged, suggesting that the system was already allocating the necessary resources to these precincts to minimize wait times. This highlights that the model was effectively addressing the needs of these priority precincts under the existing constraints.

   Recommendation: Higher minimum allocation thresholds for priority precincts should be considered when there is clear evidence of increased demand or when these areas are underserved in real-world operations. However, care should be taken to ensure such adjustments do not unnecessarily increase the total cost or resource consumption without tangible benefits.

2. **Effectiveness of Dynamic Resource Reallocation During Peak Periods**
   Introducing dynamic reallocation of resources based on predicted peak demand periods resulted in a slight increase in overall queue lengths, but there was a potential improvement in service during high-demand times. While the overall waiting time across the system increased marginally, this strategy ensured that resources were deployed where they were most needed during peak hours. This trade-off suggests that dynamic adjustments could help manage fluctuating demand more effectively, especially in precincts with variable voter turnout.

   Recommendation: Implement real-time adjustments based on predicted or observed peak voting periods. The system should incorporate adaptive scheduling and allocation strategies that can be fine-tuned based on live data, ensuring more responsive and equitable service for voters during times of high demand.

3. **Sensitivity to Voter Turnout Predictions**
   Changes in voter turnout, especially in priority precincts, significantly impacted the optimal allocation of check-in stations and voting booths. A reduction in voter turnout led to more efficient use of resources, resulting in reduced queue lengths. This emphasizes the importance of accurate voter turnout predictions to optimize resource allocation. A 10% error in predictions can drastically alter the system's performance, with increased waiting times if turnout is overestimated.

   Recommendation: Invest in refining turnout prediction models and conduct sensitivity analyses to understand the potential variability in voter turnout. This will help in adjusting resource allocations dynamically and preparing for worst-case scenarios, reducing the risk of over- or under-allocation.

4. **Doubling Check-In Stations to Reduce Wait Times**
   Doubling the number of check-in stations proved to be one of the most effective strategies for reducing overall wait times. The objective value decreased by 23.5%, indicating that the check-in stage is a major bottleneck in the voting process. Since check-in stations require staffing and can often be slower due to administrative tasks, increasing their number results in a more efficient flow of voters through the process.

   Recommendation: Focus on increasing check-in capacity, as this is the primary factor driving wait times. If additional resources are available, investing in check-in stations should be prioritized over expanding voting booths, especially in precincts where check-in delays are the most prominent.

5. **Stochastic Modeling for Arrival Rates and Service Times**
   The current model assumes static arrival rates and service times, which may not fully reflect the variability seen in real-world voting environments. Voter arrival patterns can fluctuate throughout the day, and service times can be affected by factors such as voter familiarity with the process or staffing efficiency. Modeling these parameters as stochastic variables would allow the system to better reflect real-world dynamics and improve resource allocation under uncertain conditions.

   Recommendation: Incorporate stochastic processes to model arrival rates and service times, enabling the system to handle variability and adapt in real time. This would lead to more robust performance under different voting conditions and improve the reliability of predictions for queue lengths and wait times.

Appendix

- OptiGuide transcript included in Analysis.
- Data for Census gathered from: https://data.census.gov/
- Data for Elections gathered from: https://alleghenycounty.us/Government/Elections/Election-Results
- Data for Shapefiles gathered from: https://catalog.data.gov/dataset/tiger-line-shapefile-current-state-pennsylvania-census-tract