Predicting the county-level prevalence of Chlamydia in the United States

Caroline Chocholak
University of Southern California
Viterbi School of Engineering
chochola@usc.edu

Gwenyth Portillo-Wightman
University of Southern California
Viterbi School of Engineering
gportill@usc.edu

Gabrielle (Ella) Roberts
University of Southern California
Viterbi School of Engineering
gbrobert@usc.edu

Valerie Wang
University of Southern California
Viterbi School of Engineering
wangvale@usc.edu

Abstract

The United States is currently experiencing increasing rates of sexually transmitted diseases (STDs). Rates of contraction are at record highs, while the budgets for prevention programs are being cut. By identifying which areas are at risk for increased infection rate, policy makers can target efforts towards prevention and treatment (specifically the introduction of new STD clinics) in areas that have the highest occurrence of STDs and lowest access to clinics. In this project, we develop a spatiotemporal machine learning model to predict the prevalence of chlamydia on a county-by-county level using data from the Center for Disease Control and census data. In addition, we use mixed integer linear programming to optimize the placement of new STD clinics at the county level. We analyze the locations of current STD clinics in Illinois and New Jersey to propose where STD treatment centers would optimally be placed in the future.

1 Introduction

In the United States, one in two sexually active people will contract a Sexually Transmitted Disease (STD) by age 25 [12]. In addition, while people between the ages of 15 to 24 make up the majority of STD cases in the United States, only 12% of those individuals report being tested for STDs in the last year [12]. According to the Center for Disease Control and Prevention (CDC), STD rates are at a record high for the fourth year in a row [2], making it clear that STDs are a problem that is only increasing in urgency. A visualization of the rising STD rates can be seen in Figures 1 and 2, which show the difference in cases of chlamydia per 100,000 people in 2005 and 2015, respectively. The number of red-shaded counties has increased, showing the greater prevalence of chlamydia over time.

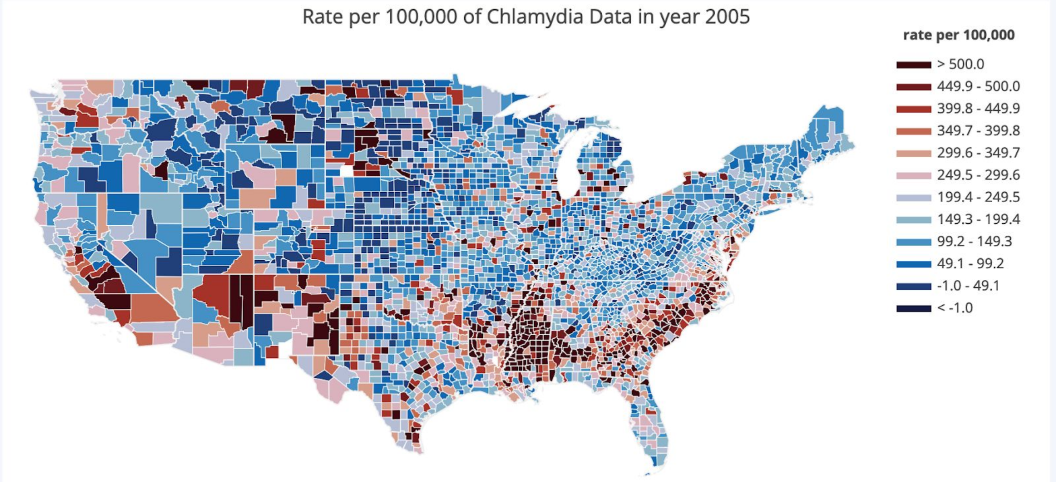


Figure 1: Cases of Chlamydia per 100,000 People in 2005

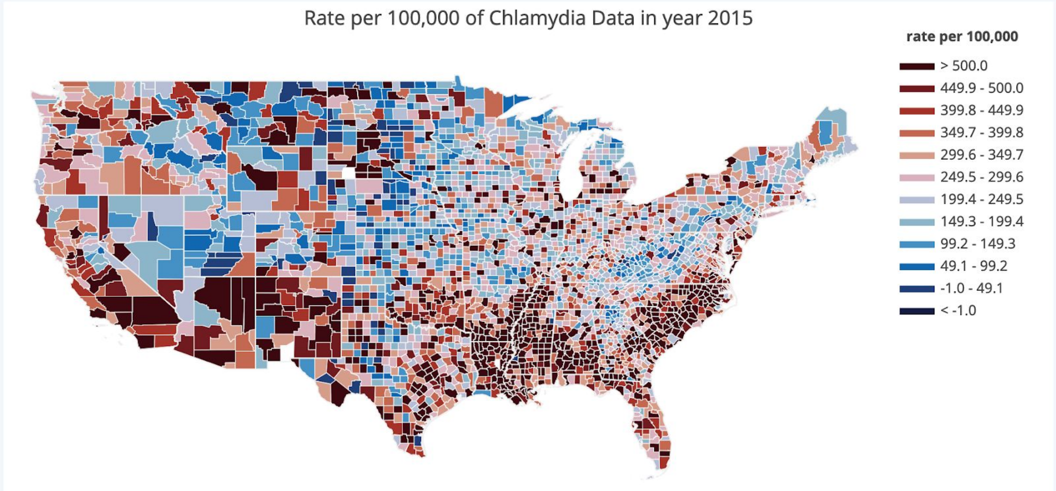


Figure 2: Cases of Chlamydia per 100,000 People in 2015

Unfortunately, there have been budget cuts in state and local STD prevention programs. In 2012, 52% of programs saw budget cuts that reduced clinic hours, contact tracing, and screening for STDs [3]. Given these budget cuts, policy makers must carefully consider where they choose to target preventative measures in order to aid the populations that are in the greatest need and to reduce the future prevalence of STDs.

One way to decrease the prevalence of STDs is to increase the number of STD testing and treatment centers around the United States. However, the number of STDs cases varies by geographic region, and the Department of Health has a limited budget to allot toward treatment centers. Considering these constraints, the placement of STDs treatment centers should be optimized for treatment needs, which requires a method to determine which areas experience the greatest occurrence of STDS and are therefore in the greatest need for STD clinics. In order to enable health committees and policymakers to plan for new STD clinics, this problem requires predicting, based on data about the past occurrence of STDs, the future rates of STD incidence. This paper centers on developing a predictive model that accomplishes this and solving an optimization problem to optimize the placement of STD clinics.

In this project, we focus on the occurence of one STD, Chlamydia trachomatis, within the United States. First, we present a predictive model that determines the number of cases of chlamydia per person at the county level, using STD, census, and county-to-county migration data. This spatiotemporal model uses data for five years and then predicts chlamydia prevalence for the sixth year. It can be used to pinpoint the areas with the highest chlamydia occurrence rates, suggesting where STD treatment centers may be most necessary in the future. Secondly, we examine the placement of new STD centers as an optimization problem, using data about the current locations of treatment centers in Illinois and New Jersey and the predictions made by the predictive model. With this data, we then use mixed integer linear programming to maximize the number of people who have access to treatment centers and to ensure that regions with high rates of chlamydia have access to an STD treatment clinics. Together, the predictive model and optimization problem can be used to select the best locations for additional treatment and testing centers. We hope that this predictive model of STD prevalence and optimization solution will provide policymakers and health committees with insight on how to ensure that STD treatment needs within the United States are adequately met.

2 Related Work

Other researchers have tackled the problem of modeling diseases and other medical conditions using machine learning, though it appears that no previous research group has created a predictive, spatio-temporal model of sexually transmitted disease for the United States.

Petrova et. al (2016) used machine learning to model chlamydia over space and time in London [10]. The researchers had a similar goal to our goal for this project: to identify high risk areas in order to inform prevention efforts by public health organizations. Their methodology was divided into three stages: visualization, exploratory data analysis, and space-time modeling. The researchers first created maps of the spread of chlamydia over time in order to visualize which boroughs of London had higher rates than others. The researchers used three different predictive modeling techniques: support vector regression, ETS, and Croston's method. They examined smaller temporal and spatial scales than we did in this project, and they did not use census or demographic data.

Maharana and Nsoesie (2018) examined how features of the built environment (such as parks, highways, and crosswalks) of cities are associated with the prevalence of obesity in those cities [7]. The team used a convolutional neural network and elastic net to extract data about the features of the environment from satellite images. Although our project uses demographic features rather than features of the environment, this work is informative about how spatial information can be used to predict the prevalence of medical conditions.

Although population surveys have traditionally been used for studies that examine public health statistics, Luo et. al (2015) demonstrated the feasibility of using just the sociodemographic data from censuses and community surveys to make inferences about regional public health outcomes [6]. Their team of researchers built a machine learning model for predicting the regional health outcomes of several non-communicable diseases across the United States, using sociodemographic characteristics from the American Community Survey. Their model was able to produce reasonable predictions that were highly correlated with the sociodemographic features. This work suggests that the census features could be used to predict STD prevalence.

2.1 Improving Upon Past Work

We aim to create a predictive model for STD rates within the United States, which no previous works have attempted. Previous similar work has also used more fine-grain data, at borough and postcode level and sampled monthly and weekly. We do not have data on such a small spatial or temporal scale available to us. Thus, we seek to create a model which still performs optimally using data which is only collected per year and at the county level. In addition, previous works have focused on more complex models. We, instead, more simple models to see if a more interpretative model can still be effective for our goals. Our ultimate goal is to take results from the predictive model to determine optimal treatment center locations in the United States that would improve access to treatment for affected individuals. No previous research has focused on optimization for this problem.

3 Data

Our predictive model of STD prevalence involves census data, data about the occurrence of STDs in the United States, and data about the county-to-county flow of migrants. In addition to STD rate and migration data, the optimization problem requires data about the locations of clinics in Illinois and New Jersey.

3.1 Predictive Model Datasets

We use the American Community Survey (ACS) 1-Year Estimates, obtained form Social Explorer, as our source of census data on the county level [13]. Census variables included total population, population density, average household size, population counts for males and females, and population count breakdowns by age, marital status, poverty status, and income. The Social Explorer website offers ACS 1-Year Estimates beginning with the year 2006, therefore we selected this year as the beginning of our data set and used data until the year 2016, which is the last available year for STD data from the CDC's Sexually Transmitted Disease Surveillance Data [1]. The census data provided information on 782 counties in the United States, rather than for all counties. The counties included in the ACS 1-Year Estimates are highlighted in red in Figure 3.

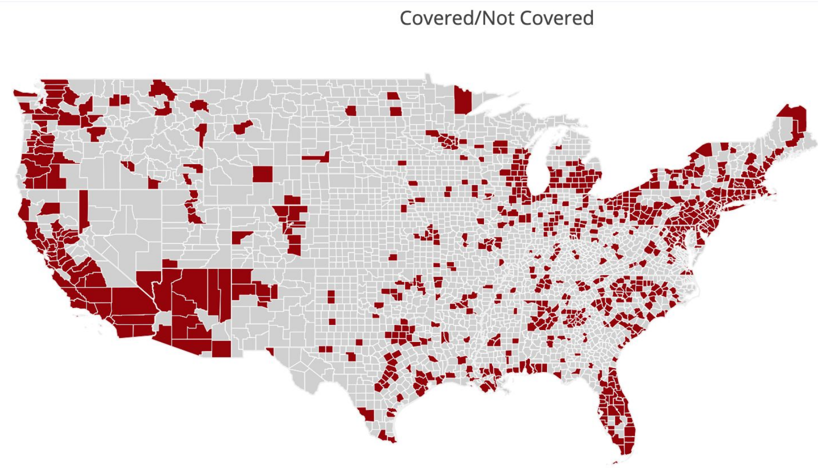


Figure 3: Map of Counties Covered (in Red) by ACS 1-Year Estimates

The CDC's STD Surveillance Data provided the number of cases of chlamydia per county for each year from from 2000 to 2016 and for counties within all states of the United States. We obtained data beginning in year 2006 and ending in 2016. The STD Surveillance Data covers all counties of the United States, but only data for 782 counties could be used, as the American Community Survey Estimates were available only for a subset of all the counties.

The SOI Tax Stats Migration Data from the IRS for the years 2006 to 2016 was used to obtain counts of the inflow and outflow of individuals between counties [4].

Finally, we used a list of counties and their neighbors from the National Bureau of Economic

Research, which lists all counties of the United States and provides the names of the counties that border a given county [8].

3.2.1 Preprocessing of Data for the Predictive Model

The CDC's STD Surveillance data is provided by county name, rather than FIPS code, therefore we preprocessed the data in order to associate each row of the STD data with a FIPS code [1]. Similarly, we had to convert the list of counties and their neighbors to FIPS codes, as they were originally provided with full county names. Additionally, all features except cases per person were normalized.

The target feature to predict was cases of chlamydia per person. The CDC provides only the raw number of cases of chlamydia per county in each year. To calculate cases per person, the number of cases per county per year was divided by the total population of the corresponding county for the given year, taken from the census data.

To capture temporal trends in the cases of chlamydia per person, our model considers the difference in the rates of cases per person for each year of training data. These engineered features involved taking the difference between the cases per person per county for the last and first years of training data, as well as between $year_1$ and $year_0$, $year_2$ and $year_1$, $year_3$ and $year_2$, and $year_4$ and $year_3$.

We aimed to capture the spatial trends in the number of cases per person by considering the flow of people between counties, calculating infected inflow *i,* which is the number of infected people from neighboring counties moving to a particular destination county *m* in each year.

For each year, we examined each destination county *m*. For each of *m*'s neighboring counties *n* (determined from the county adjacency data), we obtained the number of migrants moving from the neighbor *n* to the destination *m* in a given year, taken from the migration data. We then found the infected inflow *i*□ by summing, over all of *m*'s neighbors, the product of the number of migrants from *n* to *m* and the probability of being infected with chlamydia in county *n*.

$$i_m = \sum_{n_m} numMigrants_{m,n} \times probInfected_n$$

The probability of being infected was the number of raw cases of chlamydia in county *n* for a particular year, divided by the population of *n* in that year.

$$probInfected_n = \frac{rawCases_n}{totalPop_n}$$

Infected inflow was calculated for each county in each year.

3.2 Optimization Problem Datasets

Our optimization problem requires data on where STD treatment centers are currently located. This data is not available for the United States as a whole, therefore we selected two states, Illinois and New Jersey, that we were able to find a list of clinics for. This data came from the Illinois Department of Public Health [11] and the New Jersey Department of Health [9] and provided the name of the clinics, their street addresses, counties, zip codes, and phone numbers.

3.2.1 Preprocessing Data for Optimization Problem

The addresses of STD clinics in Illinois [11] and New Jersey [9] were preprocessed in QGIS to geocode the addresses to latitudes and longitudes so that we could more precisely visualize current locations on a map of each state. In addition, we counted the number of current clinics per county. For Illinois (which did not have census data available for all counties and therefore our predictive model could not make predictions for the state), we optimized on the summed the number of reported chlamydia cases for each county over the years 2012-2016. For New Jersey, we optimized on the summed number of cases per person (determined by the linear regression predictive model) for 2012-2016.

We attempted to compute the per-person accessibility to clinics in New Jersey and Illinois. However, since there is not ACS 1-Year estimates available for every county in Illinois, we were unable to compute the per-person accessibility for Illinois' infected population. However, we were able to determine per-person accessibility for New Jersey. In addition, we summed up the number of clinics in each county of the two states, finding that Illinois has 102 counties but only 59 STD treatment centers (six of which are within Chicago alone), whereas New Jersey has at least one STD treatment center per county.

4 Predictive Model

The machine learning methods we used to develop the predictive models were linear regression, gradient boosting regression, and random forests to predict the number of cases per person.

4.1 Time Independent Linear Regression

The first model was a time independent model, which was trained on data from all years. This model was for exploratory purposes to see how the model would work with only census and infected inflow features, ignoring temporal trends in the cases per person. This model had 43 features, which were the census features plus infected inflow. A plot of predicted and actual cases per person can be seen in Figure 4.

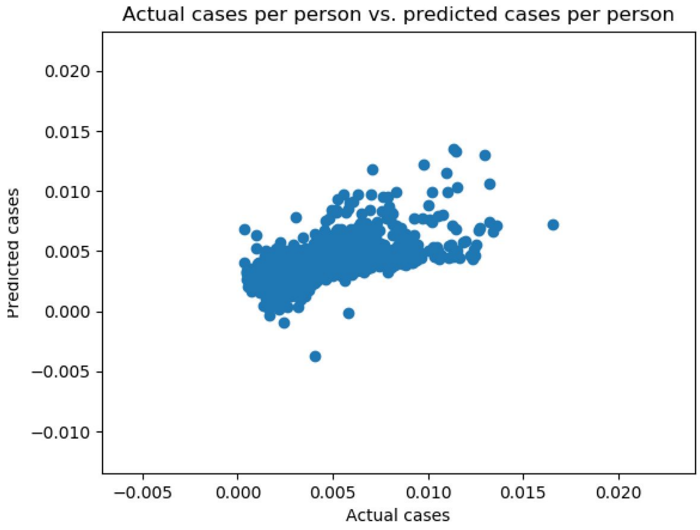


Figure 4: Predicted Cases Per Person vs. Actual Cases Per Person for a Linear Regression Time Independent Model

The r² for the model was 0.364. This low r² was an indication that census features and spatial data alone were not enough to predict the number of cases per person well. Our goal then became to develop a time dependent model that would more accurately predict the number of cases per person.

4.2 Time Dependent Models

4.2.1 Approaches to Time Dependent Models

After exploring the time independent model, we built several time dependent models, using linear regression, gradient boosting regression, and random forests, and attempted to account for the spatiotemporal trends in chlamydia rates. The gradient boosting regressor had 100 estimators, each with a maximum depth of 2. The random forest regressor had 100 estimators as well.

The time dependent models were trained on data from five consecutive years, and the model then made a prediction about the number of cases per person for counties in the sixth year. For each training year, the model took in 42 census features, infected inflow, and the number of cases per person in each training year. It also included the five engineered features that tracked the difference in number of cases per person between years, resulting in 225 features.

For each of the three machine learning methods we used, we developed six time dependent models. This was because our available data spanned from 2006 to 2016, and we split the data into 6 six-year windows so that the models could train on five years of data and predict for the sixth. The year$_0$ values for the six training models ranged from 2006 to 2011, which meant the years we predicted cases per person for ranged from 2011 to 2016.

4.2.1 Time Dependent Models Results Before Feature Selection

The models were evaluated using 5-fold cross-validation. The r^2 and MSE results of the time dependent models before feature selection can be seen in Table 1.

Table 1: R Squared and MSE Values for the Predictive Model Before Feature Selection

	Linear regression		**GBR**		**Random forest**	
Year$_0$	r^2	MSE	r^2	MSE	r^2	MSE
2006	0.707	-1.601e-06	0.907	-5.430e-07	0.905	-5.538e-07
2007	0.863	-6.819e-07	0.935	-3.195e-07	0.931	-3.432e-07
2008	0.792	-8.729e-07	0.905	-3.952e-07	0.905	-3.957e-07
2009	0.755	-1.066e-06	0.874	-5.442e-07	0.888	-4.864e-07
2010	0.746	-1.061e-06	0.894	-4.436e-07	0.893	-4.476e-07
2011	0.482	-4.347e-06	0.642	-3.501e-06	0.685	-3.307e-06

All three machine learning methods performed the best, in terms of r^2, on year$_0$ = 2007. They also all performed poorest, in terms of r^2, on year$_0$ = 2011. GBR and random forests performed much better overall, with r^2 values generally around 0.9, while linear regression performed with r^2 values around 0.7 to 0.8 (ignoring 2011, which had much worse performance across the different kinds of models). Figures 5, 6, and 7 display the actual and predicted cases per person for year$_0$ = 2007 using linear regression, gradient boosting regression, and random forests.

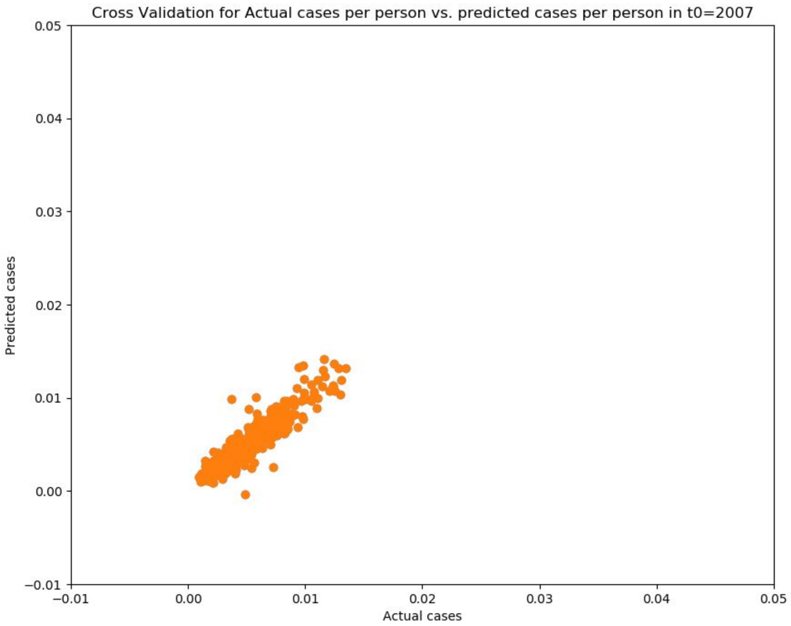


Figure 5: Predicted Cases Per Person vs. Actual Cases Per Person Using Linear Regression Time Dependent Model, $year_0$ = 2007

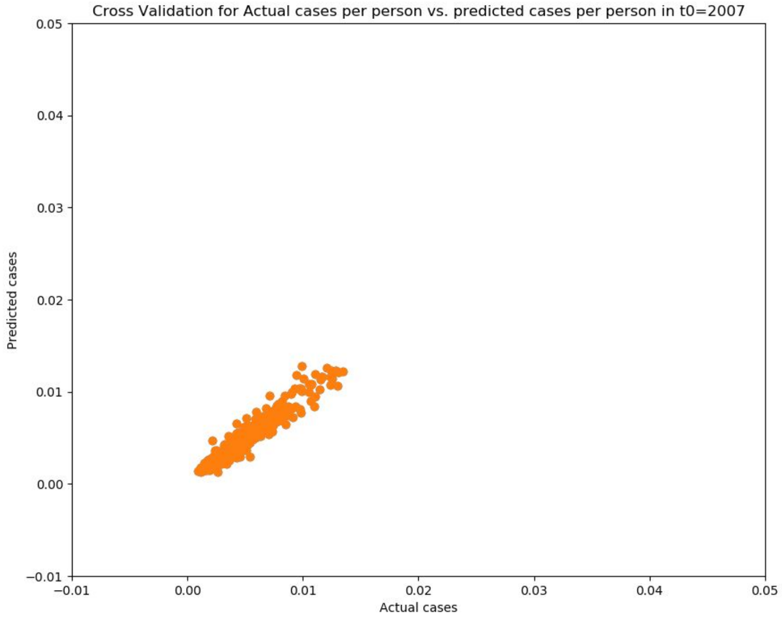


Figure 6: Predicted Cases Per Person vs. Actual Cases Per Person using GBR Time Dependent Model, $year_0$ = 2007

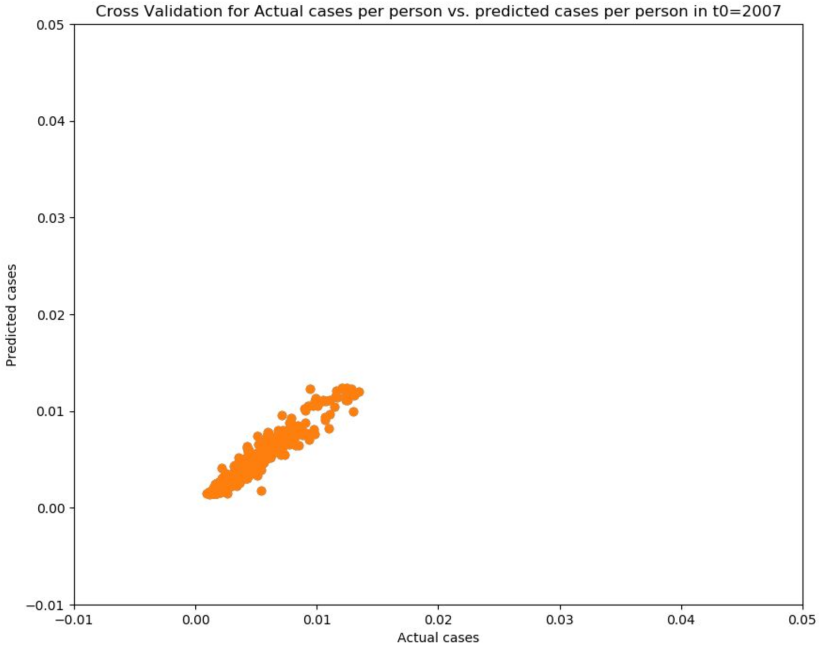


Figure 7: Predicted Cases Per Person vs. Actual Cases Per Person using Random Forest Time Dependent Model, year$_0$ = 2007

The improvement of GBR and random forests over linear regression is especially clear for the worst-performing year across the models: year$_0$ = 2011. For year$_0$ = 2011, linear regression had an r^2 of 0.482, while GBR and random forests improved that r^2 by about 0.2. Figures 8 and 9 show the difference in performance between linear regression and GBR for year$_0$ = 2011. The GBR model clearly produces more linear results.

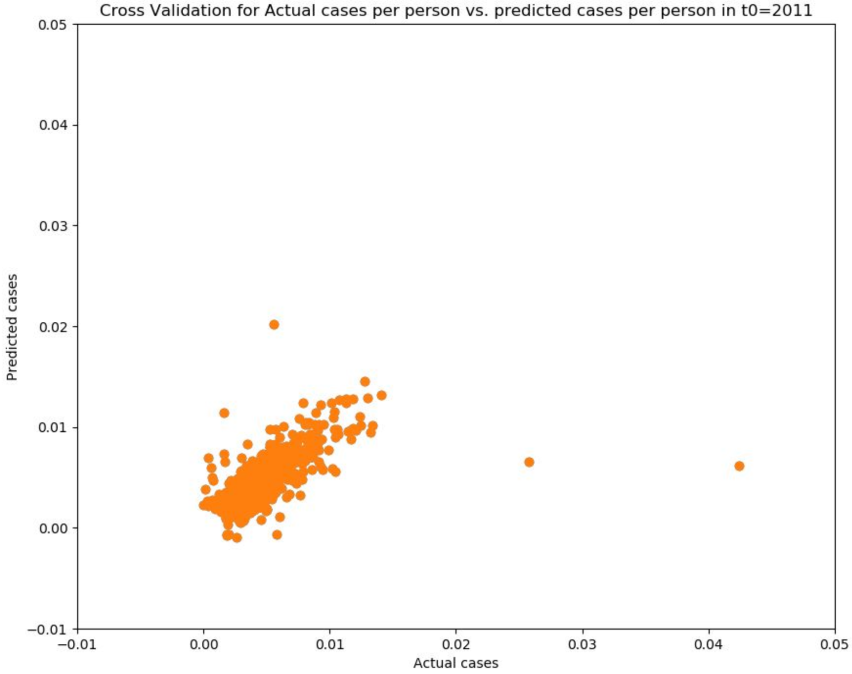


Figure 8: Predicted Cases Per Person vs. Actual Cases Per Person using Linear Regression Time Dependent Model, year_0 = 2011

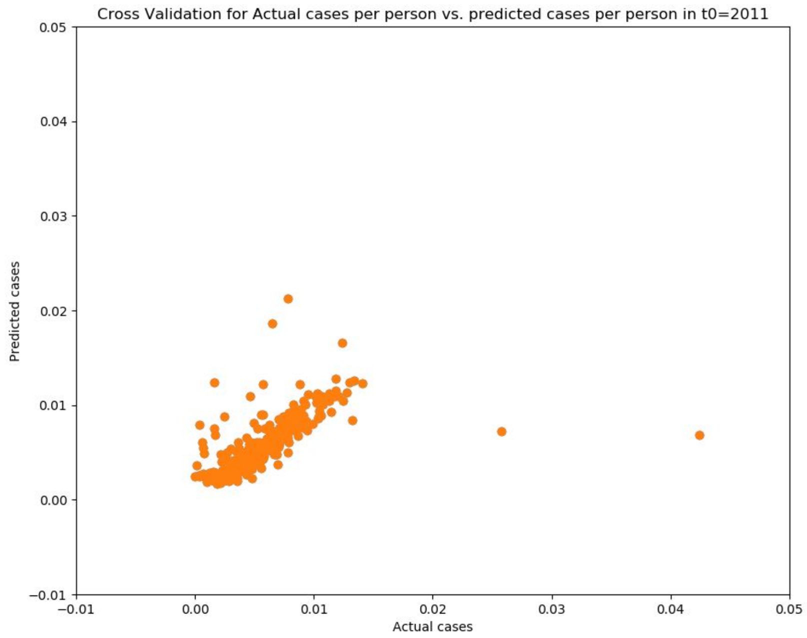


Figure 9: Predicted Cases Per Person vs. Actual Cases Per Person using GBR Time Dependent Model, year_0 = 2011

There are a few outliers that bring down the overall performance of both linear regression and GBR for year_0 = 2011, resulting in a lower r^2 than other year_0 values. For 2011, the actual number of cases was much higher than the predicted number of cases. This suggests that the models may have limited utility in the real world if health policy makers were using them to predict the prevalence of STDs because they underpredict for certain outlier areas.

4.2.2 Time Dependent Models Results After Feature Selection

We used recursive feature elimination with 5-fold cross validation to determine the 20 most important features for each time dependent model. One example of feature importance from recursive feature elimination can be seen in Figure 10, for the linear regression model of year_0 = 2006, showing that it has the best performance with about the top 20 features.

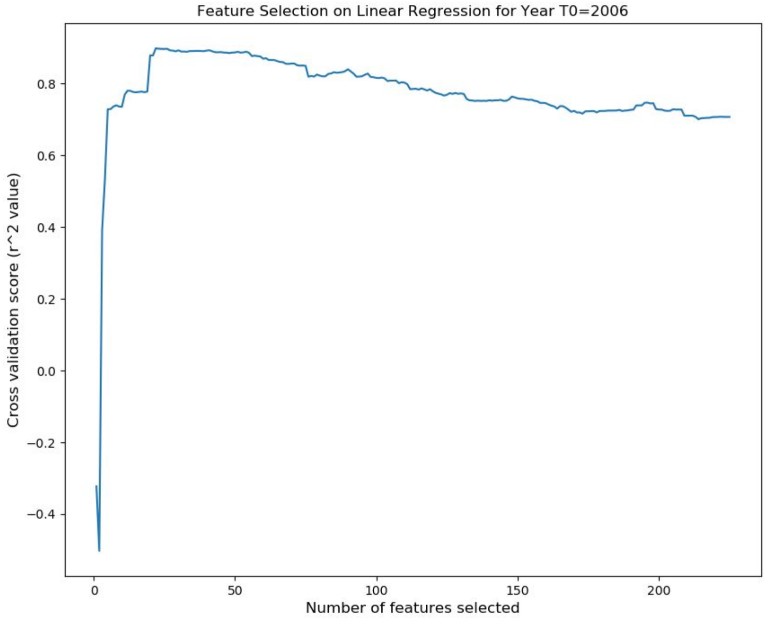


Figure 10: Recursive Feature Elimination for Linear Regression Model for year$_0$ = 2006

The r^2 results of the time dependent models after feature selection for the top 20 features can be seen in Table 2.

Table 2: R Squared Values for Predictive Models after Feature Selection

Year$_0$	**Linear regression**	**GBR**	**Random forest**
2006	0.856	0.878	0.8823
2007	0.913	0.929	0.925
2008	0.889	0.899	0.887
2009	0.724	0.753	0.748
2010	0.917	0.924	0.921
2011	0.547	0.139	0.437

After selecting for the top 20 features, linear regression improved its performance for all year$_0$ values. The most significant improvement was for year$_0$ = 2010, where r^2 increased from 0.746 before feature selection to 0.917 after feature selection.

The GBR model generally decreased in performance with feature selection, especially for year$_0$ = 2011, where r^2 decreased from 0.643 before feature selection to 0.139 after feature selection. Figure 11 shows GBR on year$_0$ = 2011 after feature selection, which can be contrasted with its performance before feature selection in Figure 9. It may be possible that with more careful feature selection (rather than just the top 20 features), GBR may be able to maintain or improve its performance once feature selection is added.

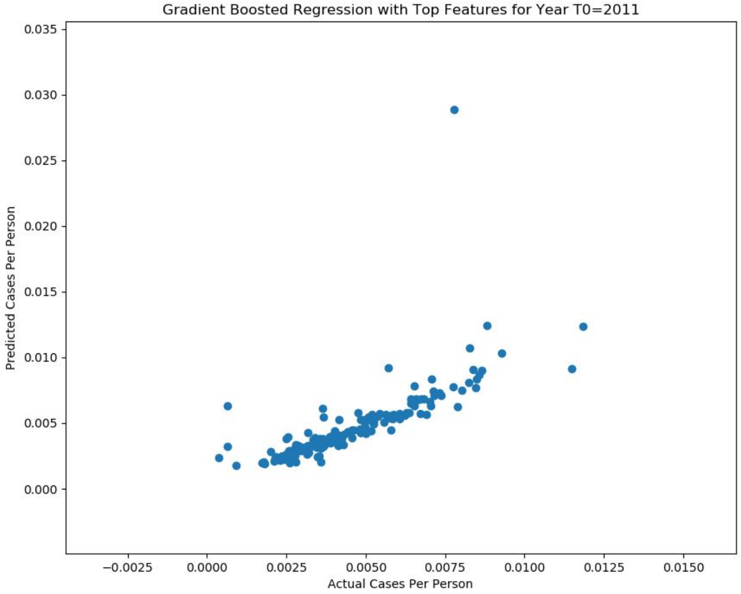


Figure 11: Predicted Cases Per Person vs. Actual Cases Per Person using GBR Time Dependent Model, $year_0$ = 2011, after Feature Selection

The performance of the random forest was also slightly worse after feature selection. The random forest for $year_0$ = 2011 was not as negatively impacted as it was for GBR. With a random forest, r^2 performance went from 0.685 to 0.437 for 2011, which is a large decrease but not as extreme as the decrease that GBR experienced.

Interestingly, in contrast to the effects of feature selection on GBR and the random forest, feature selection helped linear regression for $year_0$ = 2011, bringing r^2 up from 0.482 to 0.547.

The top features varied for each kind of machine learning technique, but they were typically fairly consistent across the $year_0$ values for a particular machine learning technique. The six most important features for each time dependent model can be seen in Table 3. This table provides only a brief overview of the trends on the most important features for each kind of model, since many of the most important features for a particular model in a certain year tended to be the same feature, just at different training years.

Table 3: 6 Most Important Features for Time Dependent Models

$Year_0$	**Linear regression**	**GBR**	**Random forest**
2006	pop_density_t2, pop_density_t0, pop_density_t3, pop_density_t1, male_t0, cases_per_person_t4	cases_per_person_t4, cases_per_person_t3, cases_per_person_t1, cases_per_person_t0, cases_per_person_t2, diff_cases_t2_t3	cases_per_person_t4, cases_per_person_t3, cases_per_person_t2, cases_per_person_t1, cases_per_person_t0, diff_cases_t3_t4
2007	pop_density_t1, pop_density_t2,	cases_per_person_t4,	cases_per_person_t4,

	pop_density_t0, cases_per_person_t4, cases_per_person_t0, cases_per_person_t3	cases_per_person_t3, cases_per_person_t1, cases_per_person_t0, cases_per_person_t2, household_income_over_200_t2	cases_per_person_t2, cases_per_person_t3, cases_per_person_t1, cases_per_person_t0, diff_cases_t3_t4
2008	pop_density_t1, pop_density_t0, cases_per_person_t4, female_t1, male_t3, female_t3	cases_per_person_t4, cases_per_person_t1, cases_per_person_t1, cases_per_person_t3, cases_per_person_t0, infected_inflow_t3	cases_per_person_t4, cases_per_person_t2, cases_per_person_t3, cases_per_person_t1, cases_per_person_t0, infected_inflow_t3
2009	cases_per_person_t4, cases_per_person_t3, age_55_to_64_t1, female_t0, cases_per_person_t2, male_t3	cases_per_person_t3, cases_per_person_t4, cases_per_person_t2, cases_per_person_t1, cases_per_person_t0, avg_household_size_t2	cases_per_person_t3, cases_per_person_t4, cases_per_person_t2, cases_per_person_t1, cases_per_person_t0, diff_cases_t0_t1
2010	cases_per_person_t4, cases_per_person_t3, male_t0, diff_cases_t2_t3, diff_cases_t0_t4, cases_per_person_t0	cases_per_person_t4, cases_per_person_t3, cases_per_person_t1, cases_per_person_t2, diff_cases_t3_t4, cases_per_person_t0	cases_per_person_t4, cases_per_person_t3, cases_per_person_t2, cases_per_person_t2, cases_per_person_t0, diff_cases_t2_t3
2011	age_25_to_34_t1, male_t2, female_t0, age_under5_t0, age_25_to_34_t0, age_under5_t1	cases_per_person_t3, household_income_150_to_200_t1, cases_per_person_t4, cases_per-person_t1, cases_per_person_t0, cases_per_person_t2	cases_per_person_t3, cases_per_person_t4, household_income_150_to_200, cases_per_person_t1, cases_per_peron_t0, cases_per_person_t2

Overall, in linear regression, the top 20 features usually included population density, cases per person at all of the training years, and male and female populations at most training years. The most important features for GBR were cases per person at all years, the difference in cases per person between years, average household size, infected inflow, and various income variables (including $75,000-$100,000, $100,000-$125,000, and 200,000+). Random forests had a similar set of important features, but income features were not as high ranking in the top 20 features. Overall, the linear regression model did not use as many of the spatiotemporal features (like infected inflow and difference in cases) as GBR and random forests did.

4.2.3 Time Dependent Models Discussion

GBR and random forest clearly perform better than linear regression in predicting cases per person. In determining which model performs best, it is important to consider how interpretable the models are. In general, linear regression has the advantage of being easily interpretable. However, given its suboptimal performance in comparison to the other methods, GBR or random forests may be a better option than linear regression for this problem.

Despite being less interpretable than linear regression, GBR still is fairly interpretable, especially because the weaker learners built in a gradient boosting regressor are very simple. The gradient boosting regressors in this project were built up of 100 weak learners, each of which had a maximum depth of 2. This low maximum depth makes the regressor still fairly interpretable overall. Given the strong performance and interpretability of GBR, we consider the GBR to be the strongest method for this problem.

4.3 Improved Time Dependent Model, Using Training Sets of the Previous Year to Predict the Next Year

To see if our linear regression model would improve and to eliminate bias with the training/test setup, we then trained a time dependent model on data for all counties in the previous year and predicted for the next year. For example, to build a model for 2008, we trained the model on five consecutive years of data, 2002 to 2006, plus the target value cases per person for the sixth year, 2007. The model then predicted predicted the cases per person for 2008 after receiving feature data about years 2003 to 2007.

This approach was done in order to possibly improve the results of previous time dependent models. Since the previous training and test sets were being split between one year's data, the previous models were training on a subset of the counties and did not have any information about the counties in the test set that were being predicted. By training on a full set of data that included feature and target values for all counties, we hoped that our results would improve since the model would now have information on all of the counties it was trying to predict.

4.3.1 Results with Improved Linear Regression Time Dependent Model

We ran our linear regression with our modified training set and with feature selection. Each year only ran with the top 9 features because we saw an overall trend that nine features showed a generally optimized result. The popular features were similar with the previous trials. Table 4 compares the results of linear regression in previous time dependent model and in the improved time dependent model.

Table 4: r^2 Comparisons For Different Training Sets for Linear Regression Model

Year To Predict	**r^2 for Train/Test on Same Year**	**r^2 for Train on Previous Year**
2007	0.913	0.863
2008	0.889	0.904
2009	0.724	0.869
2010	0.917	0.896
2011	0.547	0.600

Compared to the training set which used the same year, the results for the training set with the different year were more similar across years except for 2011. This was expected because since the training set is always using all of the counties, the result between the years should no longer differ dramatically in the improved model. The training set using the same year differs more between each year's results due to the fact that the model is using different counties for training each time the model is trained (due to randomization of the training set) and therefore some county subsets of data allow the model to perform better than other subsets of the counties data. The year 2011 does appear to be an outlier and does not perform as well as the other years.

Overall, using a the training set with the previous year appears to be the more preferred method since variation in performance from year to year is minimized.

4.3.2 Results with Improved Gradient Boosted Regression Time Dependent Model

We also ran our model with a gradient boosted regression since this model performed the best with our previous training set type. This model also ran with the top nine features, and the features were comparable to the previous trials with the different training sets.

Overall, the GBR results with the training set for the previous year were similar to the linear regression. Table 5 shows the results for the improved gradient boosting regression model.

Table 5: r^2 Comparisons For Different Training Sets for GBR Model

Year To Predict	r^2 for Train/Test on Same Year	r^2 for Train on Previous Year
2007	0.929	0.891
2008	0.899	0.844
2009	0.753	0.853
2010	0.924	0.889
2011	0.139	0.586

The results had less variance between the years (except for 2011) compared to using a training set of a subset of the current year's counties. This showed that using a training set consisting of the previous year's data is preferred for gradient boosted regression since it minimizes difference between the performance of the models for different years, while also performing quite well.

4.3.2 Discussion of Improved Time Dependent Model for 2011

The model for 2011 (shown in Figure 12) performed significantly worse than other years with r^2 values around 0.6 versus other years where the r^2 values were around 0.86. The cause for this is unknown however the graph for 2011 shows outliers in the data where the predictions were inaccurate. There are two outliers which have large actual cases per person yet the model was predicting very low values. In addition, there are few points where the cases per person is low yet the model predicted large values. These outliers thus appear to drastically lower the r^2, whereas other years do not have outliers that are so apparent.

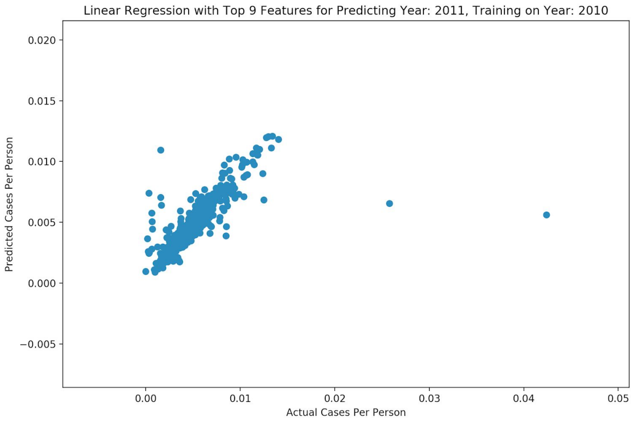


Figure 12: Improved Time Dependent Model for year_0 = 2011

5 Optimization

In addition to predicting STD rates for future years, we developed an optimization problem to determine the best counties to place STD treatment centers in order to maximize treatment to those who need it most. We formulated our problem using a Mixed Integer Linear Program that looked at current STD clinic locations (shown in Figure 13) in the specified state and the current, or predicted, infected population per county. Our optimization was multi-objective and balanced both budget and infected population's access to clinics. Due to the fact that the cost of building a clinic varies state to state, we factored in flexibility and interpretability for policymakers into our problem by allowing a varying budget (in this case, budget being the number of clinics to be built).

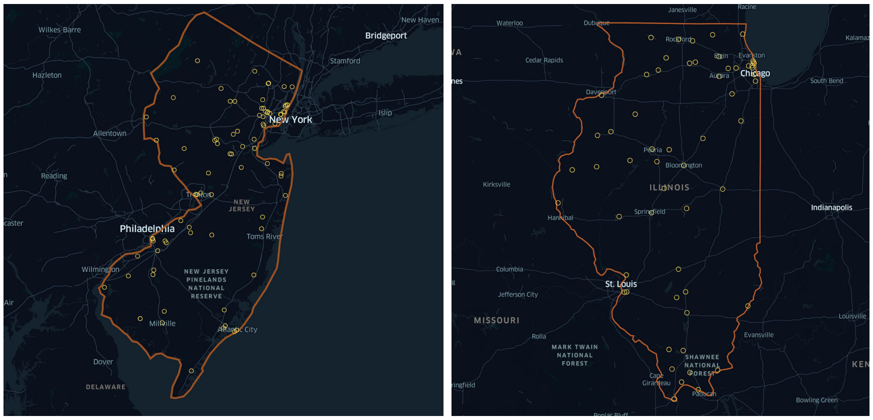


Figure 13: Maps of current STD clinic locations in New Jersey and Illinois respectively.

5.1 Mixed Integer Linear Program

Our approach for New Jersey and Illinois differed based on available data. Due to the fact that census data for Illinois did not cover all counties, we were unable to use predicted number of infected from our linear regression model and had to use the raw, current rates from the CDC. In addition, we were unable to get population counts per county and could not optimize per infected person. However, as census data covered all of New Jersey, we were able to both use both our predicted numbers from the linear regression model and optimize per person access to clinics.

We modeled our optimization off of Minimum Set Covering problem. Our decision variables, for each census tract *i*, were:

xb_i : the number of clinics built in location i
xc_i : the number of clinics covering location i

Both xb_i and xc_i must be positive integers for all census tracts *i*. Our objective was to maximize the infected population's access to clinics given a budget.

Objective = total_infected_covered – ϵ × *total_clinics_constructed*

With ϵ defined as 0.001,
With *total_infected_covered = sum*$_i$ *(*xb_i ** num_infected*$_i$*)*,
And with *total_clinics_constructed = sum*$_i$ *(*xb_i ** num_infected*$_i$*).*

That is, maximize total number of infected individuals covered and save on the number of clinics if both solutions cover the same number of people.

We also included several constraints on our problem. Coverage per county had to be less than or equal to the sum of the number of clinics in adjacent counties, plus the number of clinics in the location itself.

$$xc_i \leq \sum_i xa_i + xl_i ,$$

Where xa_i = *number of clinics in adjacent counties* and
Where xl_i = *the current number of clinics in the location itself.*

The maximum number of additional clinics built per location can be no more than two to ensure that new clinics are not all directed at one specific area with a high prevalence of STDs: $xb_i \leq 2$.

The whole problem is then constrained by the budget, which is the number of new clinics to be build, depending on the needs on policy makers.

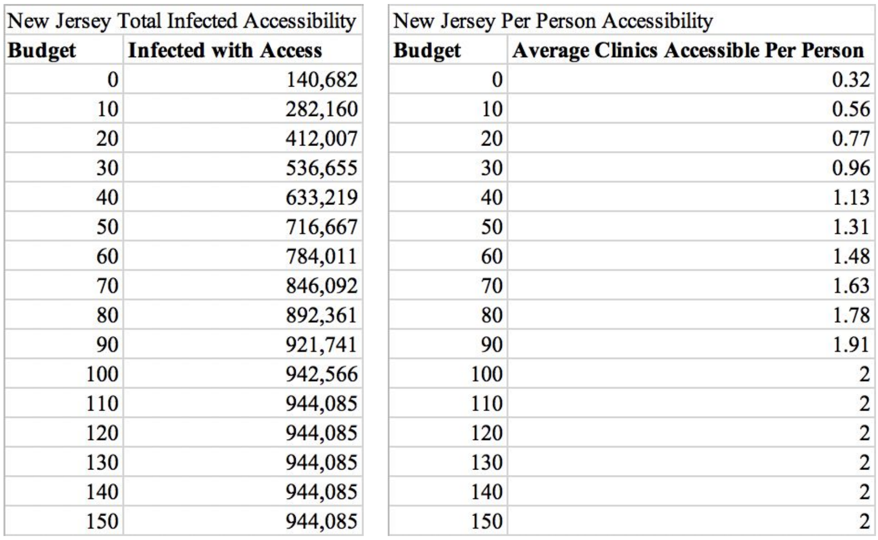
$B \geq$ total_clinics_constructed,
Where B = the number of new clinics to be built.

We programmed this formula using the Gurobi optimization solver. Since Illinois did not have an STD clinic in every county, at first, our problem only looked at coverage in a binary fashion (is a county covered or not). However, with the addition of New Jersey, where each county already had an STD clinic, we expanded our definition of coverage to include by how many clinics the county is being covered, that is the sum of how many clinics are in a particular county and in adjacent counties.

5.2 Results of Optimization

Since we created our model to incorporate flexibility, we were able to produce graphs (see Figures 14 and 15) and Tables 6 and 7 showcasing the infected population's access to STD clinics dependent on varying budgets (number of added clinics). For both Illinois and New Jersey, we found there was not much improvement in the infected population's access once 100 new clinics had been added. Illinois, however, had a much greater increase in accessibility between adding 0 to 20 clinics. We attribute this to the fact that Illinois did not have STD clinics in every county, whereas New Jersey already had at least one STD clinic in each county.

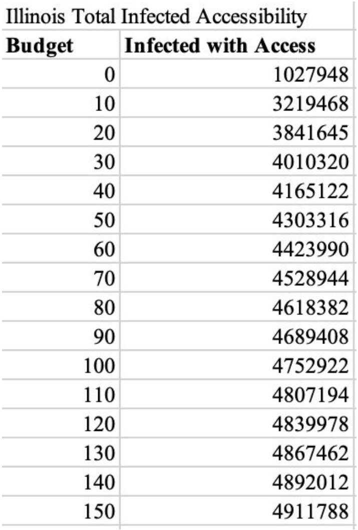
Table 6 : Charts of New Jersey's Infected Accessibility By Budget



New Jersey Total Infected Accessibility	
Budget	**Infected with Access**
0	140,682
10	282,160
20	412,007
30	536,655
40	633,219
50	716,667
60	784,011
70	846,092
80	892,361
90	921,741
100	942,566
110	944,085
120	944,085
130	944,085
140	944,085
150	944,085

New Jersey Per Person Accessibility	
Budget	**Average Clinics Accessible Per Person**
0	0.32
10	0.56
20	0.77
30	0.96
40	1.13
50	1.31
60	1.48
70	1.63
80	1.78
90	1.91
100	2
110	2
120	2
130	2
140	2
150	2

Table 7 : Illinois' Total Infected Accessibility By Budget



Illinois Total Infected Accessibility	
Budget	**Infected with Access**
0	1027948
10	3219468
20	3841645
30	4010320
40	4165122
50	4303316
60	4423990
70	4528944
80	4618382
90	4689408
100	4752922
110	4807194
120	4839978
130	4867462
140	4892012
150	4911788

Without adding any STD clinics to the state, an estimated 140,682 predicted infected persons have access to clinics within New Jersey. By adding 20 clinics, 412,007 infected persons have access. With 50 clinics, 716,667 infected people have access, or, on average, each infected person has access to 1.31 clinics. In Illinois, without adding any clinics 1,027,948 infected persons have access. By adding 20 new clinics, 3,841,645 infected people have access, an over 273% increase (compared to New Jersey's 192%).

Figure 14: New Jersey Optimization Graph: Total Infected Accessibility

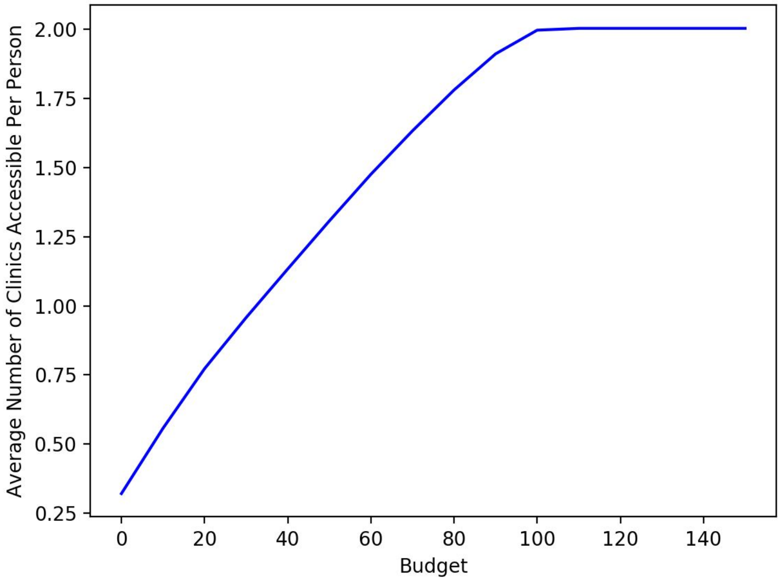


Figure 15: New Jersey Optimization Graph: Per Person Accessibility

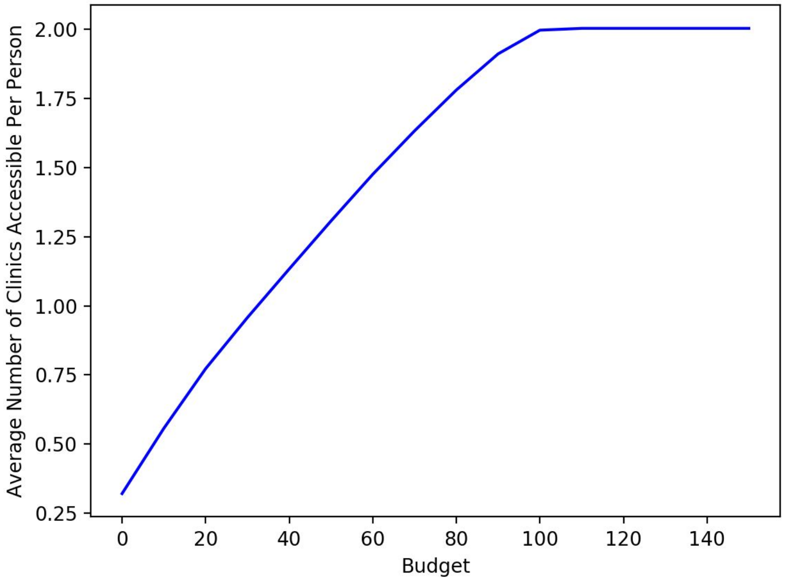
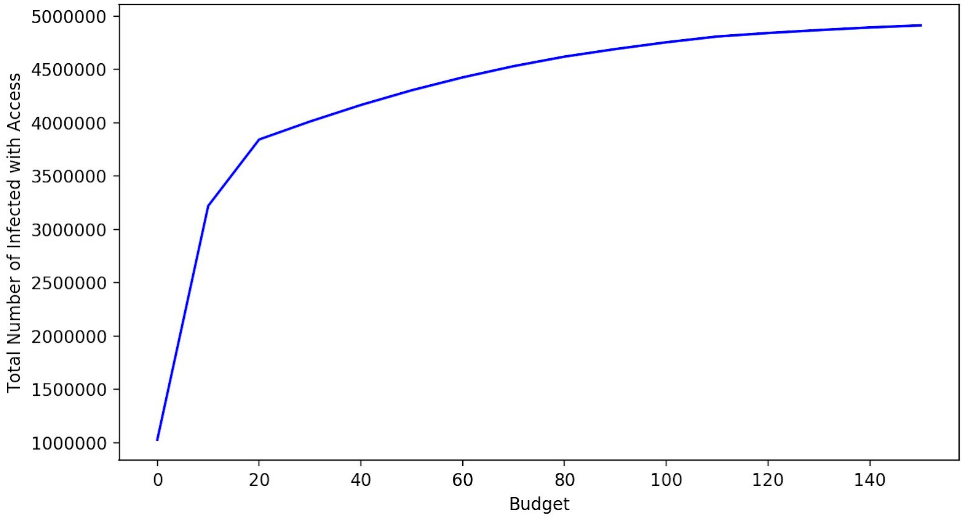


Figure 16: Illinois Optimization Graph: Total Infected Accessibility



With our optimization formula, policymakers can see in which counties to most effectively place new STD clinics. Our graph and table visualizations of number of infected access per budget also easily allow policymakers to understand the effectiveness of each additional clinic and how far their money may go.

6 Conclusions and Possible Extensions

6.1 Conclusions

Our time dependent models using linear regression, gradient boosting, and random forests all improved on the exploratory time independent model. The gradient boosting regressor and random forest performed better than the linear regression time dependent model, while developing an improved time dependent model improved on our first version of the time dependent model. Ultimately, the gradient boosting regressor may be preferred because it both performs well across all values for year$_0$ and has a high interpretability.

In the optimization problem, we found that, in both states, adding more than 100 clinics no longer increases the number of people that have access to a clinic. We also found that the number of people that gain access to a clinic with each additional clinic varies between states. This may be due to the different baselines for the states, as New Jersey began with at least one clinic per county, while some counties of Illinois did not already have a clinic.

6.2 Limitations of the Predictive Model

Our predictive model had a few key limitations, including limited data, features, and spatial and temporal resolutions.

6.2.1 Limited Data

There was only limited data available for some of the data sets used. The census data only had information for 782 counties per year, when there are actually more than 3,100 counties in total. Our model may not account for the large number of counties that are missing information, especially for states in the Midwest that may be very sparsely covered by the American Community Survey 1-Year Estimates.

6.2.2 Limited Features

We were limited in which census features we could use in our model, due to the missing values for many counties on features like school dropout rates, race, employment status, unemployment rate, and educational attainment.

6.2.3 Limited Spatial and Temporal Resolutions

Our model is limited in its utility/helpfulness for policy planners. The model predicts only one year into the future, whereas policy makers are more likely to make plans for construction and budgets many years in advance. It would also be more useful to policy makers to have finer resolution predictions, at a level lower than the county level. More ideally, the model would be able to predict at closer to the census tract level because some states have very large counties, and so having a county level prediction may not be as informative on where high STD rates occur.

6.3 Limitations of the Optimization Problem

Our formulation of the optimization problem had limitations in available data, as well as its utility and information gain for real world applications.

6.3.1 Limited Data

There is no single, unified dataset of the locations of STD clinics and treatment resources in the United States. As a result, we were only able to pick two states (New Jersey and Illinois) and find the locations of STD clinics from the state health departments. These lists of clinics from health departments may not be comprehensive of all STD clinics that are available in these areas, which could bring inaccuracy to the optimization problem.

6.3.2 Limited Utility

Our current definition of coverage is high level: a county is covered if there is a clinic in it or if an adjacent neighbor has a clinic. The clinic placement recommendations are also very high level, since recommendations are given to place a clinic in a certain county, but it is not specified where in the county. This limits the utility of the model for policy makers, as the model only provides a very general idea of where a clinic should be placed.

6.3.3 Limited Information Gain

With our current implementation of the optimization problem, we are unable to gain information about the how an additional clinic in a particular county would decrease the number of cases of chlamydia per person in that area, which would be helpful information for policy makers to have.

6.4 Possible Extensions

6.4.1 Include More Complete Data

Some future directions for the model include using more expansive, complete census data that includes all counties of the United States. Right now, the model only considers 782 of more than 3,100 counties, which is very limited. Additionally, we would like to include some of the features that we weren't able to use at this point due to missing data. This includes features like school dropout rates, race, employment status, unemployment rate, and educational attainment. It would be beneficial to incorporate not only STD clinics into the optimization problem but also STD education programs to see which areas would be benefitted by additional STD education programs.

6.4.2 Predict Further Into Future

To make the model more helpful for policy planning, we want to improve the model to make predictions for years father into the future, instead of just one year ahead. This would allow policy makers to determine the placement of new clinics ideally five years into the future, providing time for planning and construction.

6.4.3 Finer Resolution Predictions

The optimization and predictions should be at a finer resolution in the future so that more specific predictions for STD outbreaks can be made and more specific recommendations about where to put clinics can be provided. These next steps may require having STD data at finer resolution than just at the county level, since county level data is all that is available at this time.

References

[1] Surveillance and Data Management Branch. (2018). Sexually Transmitted Disease Surveillance 2017 [Online]. Retrieved from: https://www.cdc.gov/std/stats17/toc.htm

[2] Ducharme, Jamie. (2018) Americans Are Getting STDs at Record Rates. *Time*. Retrieved from: http://time.com/5379165/std-rates-sex/

[3] Howard, Jacqueline. (2018) US STD Rates Reach Record High, CDC Says. *CNN*. Retrieved from: https://www.cnn.com/2018/08/28/health/std-rates-united-states-2018-bn/index.html

[4] Internal Revenue Service. U.S. Population Migration Data [Data file]. Retrieved from: https://www.irs.gov/statistics/soi-tax-stats-migration-data

[5] Kolter, J. Z., & Ferreira, J. (2011). A large-scale study on predicting and contextualizing building energy usage. *Twenty-fifth AAAI conference on artificial intelligence*.

[6] Luo W, Nguyen T, Nichols M, Tran T, Rana S, et al. (2015) Is Demography Destiny? Application of Machine Learning Techniques to Accurately Predict Population Health Outcomes from a Minimal Demographic Dataset. PLOS ONE 10(5): e0125602.

[7] Maharana A, Nsoesie EO. (2018). Use of Deep Learning to Examine the Association of the Built Environment With Prevalence of Neighborhood Adult Obesity. *JAMA Network Open.* 2018;1(4):e181535. doi:10.1001/jamanetworkopen.2018.1535

[8] National Bureau of Economic Research. County Adjacency [Data file]. Retrieved from: https://www.nber.org/data/county-adjacency.html

[9] New Jersey Department of Health. (2016). NJ Health [PDF File]. Retrieved from: https://www.nj.gov/health/hivstdtb/documents/all_counties_clinic_sites.pdf.

[10] Petrova, D., Cheng, P.T., & Simms, D.I. (2016). Space-time modelling of sexually transmitted infections in London with focus on Chlamydia trachomatis. [PDF File]. Retrieved from: http://huckg.is/gisruk2017/GISRUK_2017_paper_97.pdf

[11] Illinois Department of Public Health. (n.d.). Sexually Transmitted Disease Clinics in Illinois [Online]. Retrieved from: http://www.idph.state.il.us/health/std/ClinicsCounty.htm

[12] American Sexual Health Association. (n.d.). Statistics. American Sexual Health Association [Online]. Retrieved from: http://www.ashasexualhealth.org/stdsstis/statistics/

[13] United States Census Bureau. American Community Survey Public Use Microdata Sample [Data file]. Retrieved from: https://www.census.gov/programs-surveys/acs/data/pums.html