

**MAKE AMERICA HEALTHY AGAIN**

Predictive and Prescriptive Analytics (F000801)

Prof. Van den Poel

Master Business engineering – Data Analytics

Academic year 2019–2020

Axelle Sels

Camille Wittouck

Thibault Verduyn

Willem Lannoye

Willem Himpe



Table of contents

[Introduction 1](#_Toc44276400)

[Exploratory analysis 1](#_Toc44276401)

[Forecasting accidents 2](#_Toc44276402)

[Methods 2](#_Toc44276403)

[Mean 2](#_Toc44276404)

[Moving average 2](#_Toc44276405)

[Exponential smoothing 2](#_Toc44276406)

[Holt linear 3](#_Toc44276407)

[Holt winter 3](#_Toc44276408)

[ARIMA 3](#_Toc44276409)

[Ensemble 3](#_Toc44276410)

[Performance 3](#_Toc44276411)

[Allocation of hospitals 4](#_Toc44276412)

[R Shiny Application 8](#_Toc44276413)

[Purpose 8](#_Toc44276414)

[Capabilities 8](#_Toc44276415)

[Conclusion 10](#_Toc44276416)

[Sensitivity of minimum population per area constraint 10](#_Toc44276417)

[Sensitivity of action radius and coverage of accidents constraints 10](#_Toc44276418)

[Time-dependent allocation 10](#_Toc44276419)

[Possibilities to expand 11](#_Toc44276420)

[R Shiny application 11](#_Toc44276421)

[Prediction improvement 11](#_Toc44276422)

[Different geographical level for the prediction 11](#_Toc44276423)

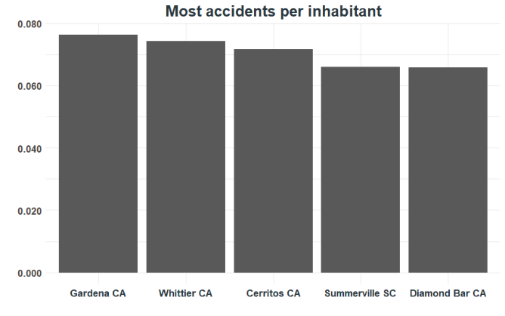
[Appendix 12](#_Toc44276424)

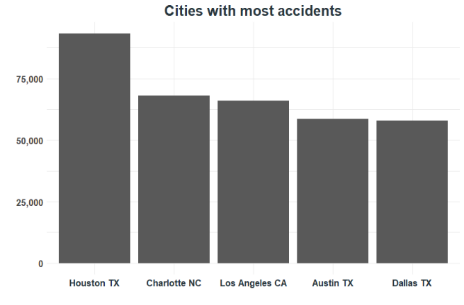
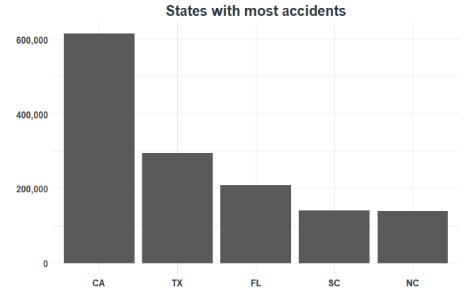
# Introduction

In order to Make America Healthy Again an easy to use tool is developed to assist with the decision making on how many hospitals to build and where the locate them. This optimization problem is based on data from car accidents over the past 4 years, since this is the most frequent cause of death. The most important parameters on which the decision to allocate hospitals is based are of course the costs and the percentage of accidents that can be reached in time.

# Exploratory analysis

The received original dataset consisted of 2,974,335 observed accidents from February 2036 to December 2039. This dataset contained a lot of features. Starting with a brief exploratory analysis, a couple of things became visible.  
Firstly, each accident gets a severity between 1 and 4. There are almost no accidents in the dataset with a severity of 1. This could be due to the fact that these accidents are not severe enough to report them. In reality, there are probably more accidents with a lower severity. The severity score of 2 is most frequent, with 67.5% of the cases receiving this score. It is however not clear what determines the severity score.   
Secondly, several POI’s are taken under scrutiny. The one that stood out was the presence (or rather absence) of speed bumps in the cases. From all the different cases, there are only 454 accidents with a speed bump present in a nearby location. This is remarkable, and could be a motive to look further into the influence of speed bumps on the occurrence of accidents.   
Thirdly, attention is given to the location of the accidents. More specifically, the states with the most accidents, the cities with the most accidents and the cities with the highest number of accidents per inhabitant were reviewed. The results can be seen below. Firstly, it is visible that the state of California has a huge number of accidents (over 600 thousand), which is more than double of the state of Texas, which is placed second in this row. Florida, Sosuth Carolina and North Carolina complete the top five. Texas however contains three of the top five cities with the most accidents, namely Houston, Austin and Dallas. Charlotte (North Carolina) and Los Angeles (California) are the other cities in these top five.





Finally, the cities with the most accidents per inhabitant were considered, however, only the cities with more than 50,000 inhabitants since this information could be useful to decide where to build the first hospitals, given that hospitals can only be built in cities with this minimum population. Four of the five cities can be found in California, namely Gardena, Whittier, Cerritos and Diamond Bar. The fifth city is Summerville (South Carolina).

# Forecasting accidents

Afbeelding met tekst, foto, kijken

Automatisch gegenereerde beschrijvingAfter the exploratory analysis, only the features regarding the location from the accident (i.e. longitude & latitude as well as city, county, state) and the time of occurrence are taken into account in order to forecast accidents. With this information, all the accidents could be aggregated per city, per county, per state and per month. The choice of forecasting per month is based on the fact that monthly data is still quite manageable, with a sufficient level of variation to capture seasonality and linear trends. However, it must be mentioned that although there is data available from February 2036, it was remarkable that certain states did not report any accidents until June 2036. Hence, there is decided to use a cut-off and use the data from July 2036 onwards, to improve the performance of our forecasting methods.

## Methods

In order to obtain the forecasts of the time series analysis, several methods were used. Each of these are briefly discussed below. Each of these have their own advantages and typical characteristics. The performance measures of all these methods can be found below, in Table 1.

### Mean

The simplest forecasting method is the mean. It takes the average of all previous observations to obtain a prediction for the next period. This method is very simple and computationally not hard at all. However, this method doesn’t take seasonality and/or linear trend into account.

### Moving average

When the mean is limited in time, i.e. it looks only at a certain number of previous periods, it is called the moving average. The moving average is similar to the mean and will lag behind if there is a certain trend or seasonality present. On top of that, when forecasting 20 years based on 4 years of observations, in the long run this method will simply converge to one single value. In this case, the last 12 months are set as the period for calculating the moving average.

### Exponential smoothing

The exponential smoothing is a good mix between impact of historic observations and recent observations. It is also possible to decide the weight of recent observations yourself. The weight is denoted by α, which is set to 0.2 in this case. As can be seen in Table 1, this method is actually the best performing in all categories and for all regions.

### Holt linear

This method is the first that is able to capture a linear trend. Holt’s linear method has 2 parameters, namely α, which is similar to the smoothing constant of exponential smoothing discussed earlier, and β which smooths the linear trend. In this case, there is decided to give both these parameters a value of 0.2.

### Holt winter

The Holt-Winter’s method builds further on the previous one. Next to a weighted average and linear trend, this method also incorporates seasonality. This implies that a third parameter γ is introduced, which takes the seasonality into account. In this case, all three parameters are set to 0.2, which is a common chosen number for these parameters.

### ARIMA

The last forecasting method that is used is the Auto-Regressive Integrated Moving Average, or ARIMA in short. This method does not use linear trends or seasonality, but is based on the fact that time series mostly contain autocorrelation. To deal with linear trend or seasonality, ARIMA needs stationary data, and if this is not the case, it differentiates the data (hence the Integrated part) to obtain stationary data. In this case, we used the auto.arima function to select the best model automatically.

### Ensemble

Finally, to create one prediction per region per month, the results of the previously discussed methods are integrated into one result. This is done by taking the weighted averages of the results of these methods. The weight is determined by the MASE. The lower the MASE, the more weight the results get. For instance, when predicting the number of accidents per state, the exponential smoothing results get a weight of 21.07% when putting it into the ensemble. In this way, the predictions can become more robust.

### Performance

*Table 1 - performance measures of forecasting methods* ***(nog updaten!)***

|  |  |  |  |
| --- | --- | --- | --- |
|  | MAE | RMSE | MASE |
| Mean\_pred\_mt\_states | 483.24 | 1562.00 | 0.2278 |
| Mov\_avg\_mt\_states | 361.74 | 1521.17 | 0.1705 |
| Exp\_smoothing\_mt\_states | 346.76 | 1532.82 | 0.1635 |
| Holt\_mt\_states | 577.55 | 1961.47 | 0.2722 |
| Holtwinter\_mt\_states | 461.63 | 1617.03 | 0.2176 |
| Pred\_arima\_mt\_states | 483.69 | 1558.18 | 0.2280 |
| Mean\_pred\_mt\_counties | 10.85 | 72.90 | 0.2383 |
| Mov\_avg\_mt\_counties | 8.60 | 68.94 | 0.1889 |
| Exp\_smoothing\_mt\_counties | 8.30 | 68.74 | 0.1823 |
| Holt\_mt\_counties | 14.48 | 100.85 | 0.3182 |
| Holtwinter\_mt\_counties | 11.28 | 74.40 | 0.2479 |
| Pred\_arima\_mt\_counties | 11.07 | 72.73 | 0.2432 |
| Mean\_pred\_mt\_cities | 2.63 | 18.84 | 0.2510 |
| Mov\_avg\_mt\_cities | 2.25 | 16.46 | 0.2151 |
| Exp\_smoothing\_mt\_cities | 2.17 | 16.15 | 0.2070 |
| Holt\_mt\_cities | 3.48 | 30.16 | 0.3317 |
| Holtwinter\_mt\_cities | 2.80 | 19.18 | 0.2668 |
| Pred\_arima\_mt\_cities | 2.61 | 21.99 | 0.2485 |

Table 1 gives an overview of all forecasting methods per region. The MAE and RMSE are closely related and give an indication of how far off the forecast is from the true values. These are of course a lot higher for states than for counties and cities. To equally compare all results and since there needed to be dealt with a lot of zero’s (especially when looking at the cities), the MASE was a great alternative to the MAPE. The test set was chosen as the last 12 months (i.e. the year 2039) while the train set were all the months before (July 2036 – December 2038). In general, the exponential smoothing performs best while Holt Winter’s methodseems to perform worst.

# Allocation of hospitals

The forecast of the number of accidents per city in the USA in the period 2040 to 2060 is made based on the historical accidents in the period July 2036 to December 2039, as mentioned before. The allocation of the hospitals will be based on the predictions from 2041 onwards, since it takes one year (2040) to build the hospitals.

To determine the exact place where to locate the hospital in a certain city, the centroids of all the historic accidents were calculated per city. Based on the coordinates of these centroids, a distance matrix is calculated from each city to each other city, considering the rows as starting point and the columns as end points, in other words a ‘from-to’ matrix. Since hospitals can only be built in cities with more than 50000 inhabitants, all the columns of cities with less inhabitants were deleted.

To effectively determine the number of hospitals and the cities where these hospitals should be built in order to minimize the cost, taking into account that 98 percent of all the accidents should be in the reach of less than 100 miles from at least one hospital, a R based genetic algorithm with binary chromosomes is used (rgba.bin). Therefore an evaluation function is written which is used for variable selection to find the optimal allocation.

This evaluation function works based on a vector with binary values that indicate if a hospital is placed in a city, with the cities in the same order as the columns of the ‘from-to’ distance matrix. The function is initiated by setting the cost to zero. Next, for each column in the distance matrix that represents a city where no hospital is allocated, the distances are put on infinity because there is no hospital in this city. Based on this adapted distance matrix that takes into account the existence of hospitals, for each city (row), the minimum distance is determined. This results in a vector that indicates how far the nearest hospital is per city. Next, this vector is transformed into a binary vector that shows a one when the maximum distance (standard 100miles) is exceeded. Next, this vector is multiplied by the number of accidents per city. In this way, it is possible to calculate the percentage of accidents that is not covered. If the minimum coverage is not reached, the cost should be set at infinity. In case that sufficient accidents are covered, the cost is simply calculated as stated in the assignment. In this way, however, the algorithm was unable to learn from its infeasible solutions, because no distinction was made between the solutions that were close to the coverage boundary, or far away. Therefore there was decided to give a huge cost that is proportional to the percentage of accidents that were not covered. This seemed to enhance learning speed of the algorithm immediately.

So, only one of the constraints is dealt with within the evaluation function, namely the minimum coverage of 98%. The constraint that concerns the minimum population is solved by making it impossible to allocate hospitals to these ‘small’ cities even before the algorithm is executed by means of the from-to matrix, as explained before.

In order to be able to adjust the various parameters in an efficient way, the evaluation function was placed in an overarching function together with the genetic algorithm. This made it possible to test different design options.

Minimum population limit: 30.000 – 40.000 – 50.000

Maximum distance to hospital: 100 – 110 – 120

Minimum coverage percentage: 0.01 – 0.02

The top two constraints were relaxed twice, while the latter one was tightened. This decision was made from an ethical perspective. The relaxation of the second constraint can be made feasible without increasing the number of victims by for instance better equipped ambulance, faster ambulance etc.

Although the size of the binary chromosome was limited to the cities that matched the population constraint, it was still very difficult to find feasible solutions. Once some feasible solutions were found for the standard settings, these were given as suggestions. This ensured that even with fewer iterations, an optimum was found.

|  |  |  |  |
| --- | --- | --- | --- |
| Design | Design features | Number of hospitals | Costs |
| 1 | (50000, 100, 0.02) | 159 | 2,65E+12 |
| 2 | (40000, 100, 0.02) | 131 | 2,42E+12 |
| 3 | (30000, 100, 0.02) | 164 | 1,62E+12 |
| 4 | (50000, 110, 0.02) | 107 | 2,42E+12 |
| 5 | (40000, 110, 0.02) | 109 | 2,11E+12 |
| 6 | (30000, 110, 0.02) | 128 | 1,23E+12 |
| 7 | (50000, 120, 0.02) | 106 | 1,49E+12 |
| 8 | (40000, 120, 0.02) | 110 | 1,54E+12 |
| 9 | (30000, 120, 0.02) | 130 | 1,21E+12 |
| 10 | (50000, 100, 0.01) | 194 | 4,04E+12 |
| 11 | (40000, 100, 0.01) | 199 | 4,38E+12 |
| 12 | (30000, 100, 0.01) | 196 | 2,56E+12 |
| 13 | (50000, 110, 0.01) | 156 | 3,33E+12 |
| 14 | (40000, 110, 0.01) | 154 | 2,23E+12 |
| 15 | (30000, 110, 0.01) | 148 | 2,49E+12 |
| 16 | (50000, 120, 0.01) | 119 | 2,54E+12 |
| 17 | (40000, 120, 0.01) | 136 | 1,84E+12 |
| 18 | (30000, 120, 0.01) | 148 | 1,61E+12 |

The table above shows how the optimal solution varies depending on the design. All the designs in this table use cities as the geographic level. (see appendix for similar table with counties as geographic level)  
It is clear to see that if the constraints are relaxed, the costs decrease. Lower costs do not necessarily mean that fewer hospitals will be built. This can be explained by the fact that the construction costs are non-recurring costs and by the fact that the operational costs depend on the number of inhabitants in the city where the hospital is located.

A possible future goal might be to reach a coverage of 99%. This leads to a 52% increase in costs if the other constraints remain unchanged. Reaching this goal while keeping the costs stable is possible when you choose for design 15 or 16, relaxing the minimum population and maximum reach constraints.

In addition, we have also investigated whether it is necessary to start building all hospitals immediately in 2040, or if it is also possible to build some hospitals at a later date. In this way, the operational costs associated with the hospitals can be reduced. We did so by adjusting the evaluation function. Also, the genetic algorithm was slightly modified. Instead of predicting a binary chromosome with the size equal to the cities that are large enough, the vector is now twenty times as long.

The chromosome is divided in twenty equal parts and the evaluation function includes a for loop that is performed twenty times. Each time the for loop is performed, a next part of the chromosome is passed. This was done by increasing the count 'year' by one unit each time at the end of the loop. In this way, each loop calculates the cost of the allocation of hospitals per year. The operational cost is calculated in the same way as above, but the building cost has become a little more complex. In year one, the building cost is simply equal to the number of ones in the first part of the chromosome. But from year two onwards, each time the difference is calculated between the part of the chromosome of that particular year and the previous year. If there is a transition from 0 to 1, building costs are charged again.

A limitation of this method is the fact that in this way the possibility exists to stop using a hospital that was in use the year before. In order to avoid this undesirable scenario, the construction costs will be charged back each time this hospital has to be made operational again.

The number of uncovered accidents is calculated for each year separately. Once the for loop has been completed, the coverage over the 20 years is calculated. So, the constraint concerning the coverage percentage is checked outside the for loop.

Even though this timeseries algorithm is executed with 100 iterations, this hasn't given the result that was hoped for. This is logically explained by the fact that our predictions for the next 20 years are only based on the 42 prior months. As a result, there is very little variation in the predictions, so the optimal scenario is to immediately build the total number of hospitals in 2040. Certainty about this finding is substantiated by the fact that this function has indeed proven its working and added value by the optimal solutions calculated with fictitious data. This fictitious data was created by giving the predictions a remarkable upward trend in the last three years. The optimal solution consequently showed a second construction period just before the increase in accidents.

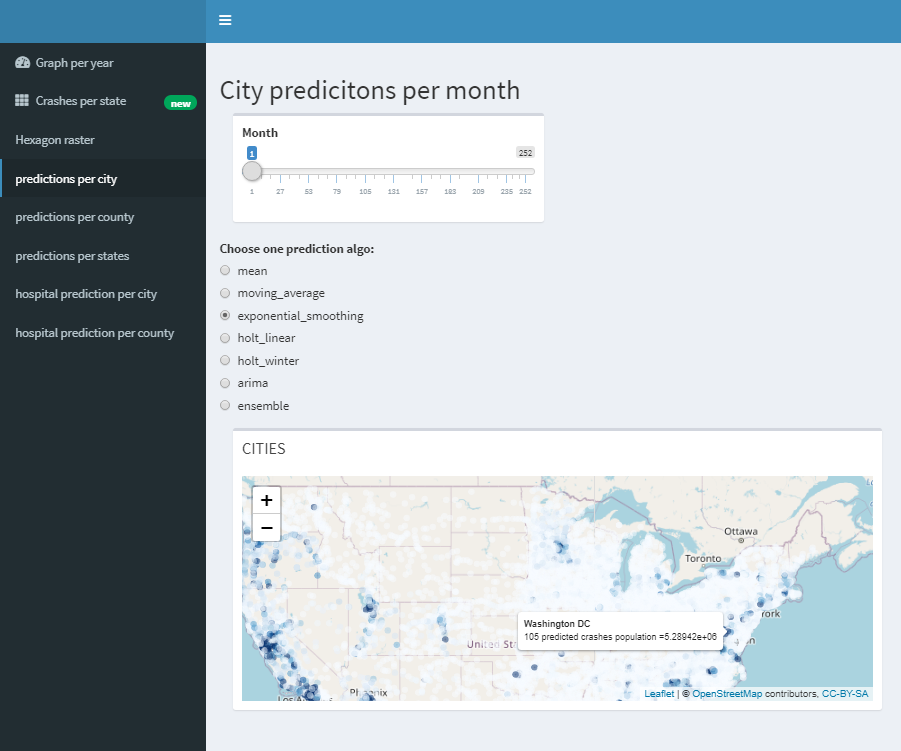
# R Shiny Application

## Purpose

This R shiny application is created for the federal government of the united states. The platform its main purpose is to bring financial and social insight in the trade-offs of the development of new hospitals across the united states. The application is built so it can be used in a mobile and desktop setting. The way the user needs to interact with the app is by changing the different settings, so the estimated costs can be given based on the entered trade-offs. The main trade-offs are the speed of arrival for the ambulances based on the distances of the hospitals. The second trade-off is the percentage of accidents that happen within a certain range of the hospital. A larger amount of hospitals can drive up the costs, but at the same time, can give the injured people a bigger chance of survival. The last trade-off is the minimum population size of the region where the hospital is located. When hospitals are placed in very dense cities, it shortens the arrival time for its local citizens, but increases its reaction time for accidents in rural parts of the country.

## Capabilities

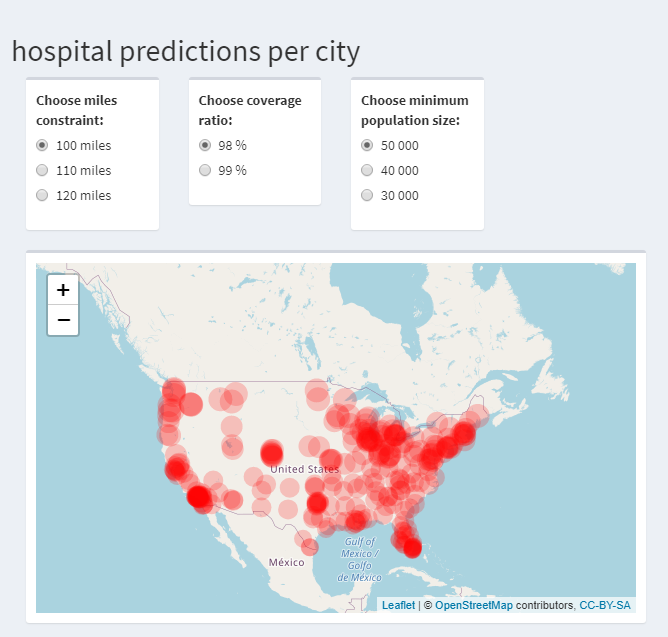
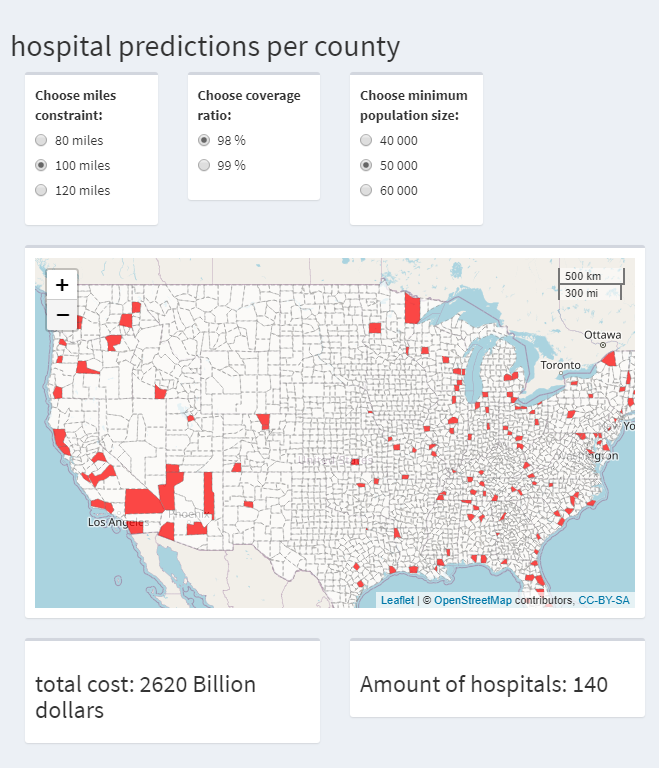
Our interactive application has descriptive, predictive and sensitive analysis capabilities. You will be able to use slider and selection buttons to set the settings that are needed. >………………

The predictive menu items are based on three different regional locations: States, Counties and Cities. For each of these regions the amount of accidents is predicted based 6 different prediction algorithms and one optimal ensemble. More information on the algorithms are in the prior paragraph. The prediction is for every month starting from January 2040 until December 2059. You can select the specific month using the top slider, and underneath you will need to select one of the 7 algorithms. When the required options are selected, you will see an interactive map where you can select a specific state, county or city. The selected region will then show the name of the region together with the amount of predicted crashes and population size.

The sensitivity analysis capabilities can be found under the hospital prediction tabs. These interactive plots are created with the predicted locations of the future hospitals using generic algorithms. More information in prior paragraphs. The user has the capability to choose from 3 different options. The first is the amount of miles that a certain percentage of the predicted accidents need to be within a hospital. The second is the coverage ratio. This ratio is the minimum percentage of accidents that need to be within the miles-constraint of a hospital. The third and last option is the minimum population size of the geographical regions where the hospital would be placed. Based on these three options, the user may optimize the hospital development plan based on financial or social incentives. The output that is given is the total cost which includes the building and operating costs but also transportation costs, and the amount of hospitals that will be built together with their location on a map. Every Marker on the interactive map has the specific radius base on the constraint that was initially given. This way the user can see where there are certain regions where there is a lot of overlap.

# Conclusion

## Sensitivity of minimum population per area constraint

For the initial constraints the optimal result shows that 159 hospitals were build when aggregating the accidents on a city level while only 104 hospitals were built when aggregating on a county level. The cost is very similar and around 2.6 trillion US dollars.   
These results also show that hospitals are typically not build in large cities or counties themselves such as LA or Brooklyn. Rather they are built in nearby counties or cities with a low population, which is of course a consequence of the high maintenance cost per inhabitant of a particular area.

## 

## Sensitivity of action radius and coverage of accidents constraints

When relaxing constraints, the costs can be lowered significantly but at the ethical cost of saving human lives. For example, when relaxing the action radius constraint (I.e. the distance from city/county to nearest hospital) the cost can be halved. It is of course the question if this is something desirable. As this would increase the average time to hospital and therefore increasing the average lethality. Ideal would be to be able to relax the action radius constraint but remaining the same chance of survival by means of for example faster and better equipped ambulances.

## Time-dependent allocation

The time component of the algorithm determines the need to build a hospital in a city, supplemented with in which year this hospital would be built. This analysis could influence the total costs, but showed that there is in fact no impact since the allocation of all hospitals occurred in year one. This is due to the fact that the predictions are relatively stable over time.

# Possibilities to expand

## R Shiny application

This application is a prototype that can still be improved based on increased variables and observations. On the visual side the possible expansions are a better selection option for prediction months. Currently, we are using a slider but a calendar where you may select the month and year might be user friendlier.

## Prediction improvement

Two possible enhancements would be the use of covariates and a longer time window.

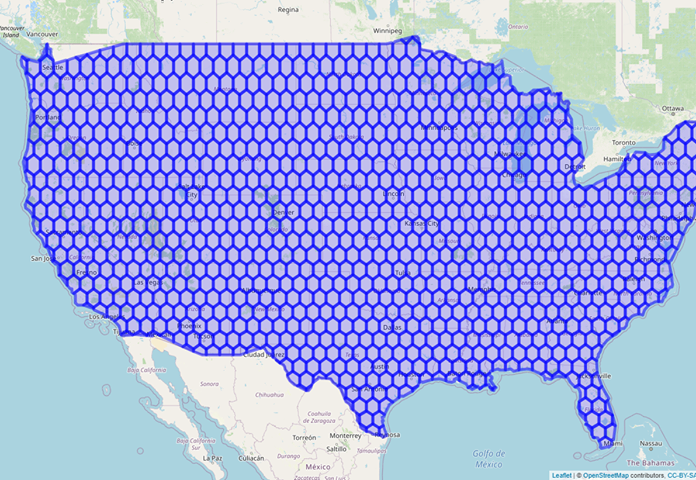
## Different geographical level for the prediction

The application now can analyze both per city and per county, but a ‘per hexagon’ level was also in the pipeline. The population of the hexagon would be approximated by the sum of the populations of cities enclosed in the hexagon.

# Appendix

Table of costs and number of hospitals dependent on different constraints using counties as geographic level.

|  |  |  |  |
| --- | --- | --- | --- |
| Design | Design features | Number of hospitals | Costs |
| 1 | (50000, 100, 0.02) | 104 | 2,62E+12 |
| 2 | (40000, 100, 0.02) | 115 | 2,42E+12 |
| 3 | (60000, 100, 0.02) | 95 | 2,98E+12 |
| 4 | (50000, 80, 0.02) | 228 | 6,13E+12 |
| 5 | (40000, 80, 0.02) | 197 | 4,98E+12 |
| 6 | (60000, 80, 0.02) | 185 | 5,44E+12 |
| 7 | (50000, 120, 0.02) | 97 | 2.03E+12 |
| 8 | (40000, 120, 0.02) | 102 | 2,21E+12 |
| 9 | (60000, 120, 0.02) | 88 | 2,11E+12 |
| 10 | (50000, 100, 0.01) | 170 | 4,38E+12 |
| 11 | (40000, 100, 0.01) | 153 | 3,77E+12 |
| 12 | (60000, 100, 0.01) | 172 | 5,15E+12 |
| 13 | (50000, 80, 0.01) | 251 | 7,17E+12 |
| 14 | (40000, 80, 0.01) | 248 | 6,66E+12 |
| 15 | (60000, 80, 0.01) | 228 | 7,64E+12 |
| 16 | (50000, 120, 0.01) | 117 | 2,81E+12 |
| 17 | (40000, 120, 0.01) | 103 | 2,52E+12 |
| 18 | (60000, 120, 0.01) | 107 | 2,78E+12 |



Map of United states divided in hexagons, due to limited time we were not able to explore this geographic level further.