Twentieth Mathematical and Statistical Modeling Workshop for Graduate Students

14 – 22 July 2014
North Carolina State University
Raleigh, NC, USA

Organizers:
Pierre Gremaud, Ilse C.F. Ipsen, Ralph C. Smith
Department of Mathematics
North Carolina State University

This report contains the proceedings of the *Industrial Mathematical and Statistical Modeling Workshop for Graduate Students*, held in the Department of Mathematics at North Carolina State University (NCSU) in Raleigh, North Carolina, 14 – 22 July 2014.

This was the twentieth such workshop at NCSU. It brought together 31 graduate students from Mathematics and Statistics Departments at 28 different universities.

The goal of the IMSM workshop is to expose mathematics and statistics students from around the country to: real-world problems from industry and government laboratories; interdisciplinary research involving mathematical, statistical and modeling components; as well as experience in a team approach to problem solving.

On the morning of the first day, industrial and government scientists presented six research problems. Each presenter, together with a specially selected faculty mentor, then guided teams of 5– 7 students and helped them to discover a solution. In contrast to neat, well-posed academic exercises that are typically found in coursework or textbooks, the workshop problems are challenging real world problems that require the varied expertise and fresh insights of the group for their formulation, solution and interpretation. Each

group spent the first eight days of the workshop investigating their project and reported their findings in 20 minute public seminars on the final day of the workshop.

The IMSM workshops have been highly successful for the students as well as the presenters and faculty mentors. Often projects lead to new research results and publications. The projects can also serve as a catalyst for future collaborations between project presenter and faculty mentor. More information can be found at `http://www.samsi.info/IMSM14`

Sponsors

Statistical and Applied Mathematical Sciences Institute (SAMSI)
Center for Research in Scientific Computation (CRSC)
Department of Mathematics, North Carolina State University

Problem presenters

Agustin Calatroni, Russ Helms, and Herman Mitchell, Rho Inc.
Matthew Farthing, US Army Corps of Engineers
Simone Gray, Centers for Disease Control & Prevention
John Peach, MIT Lincoln Laboratory
Mark Wolff, SAS Institute Inc.

Faculty mentors

Howard Chang, Emory University
Lea Jenkins, Clemson University
Kenny Lopiano, SAMSI and Duke University
Minh Pham, SAMSI
Sanvesh Srivastava, SAMSI

Projects

The Hunt for Red Hot Rock-tober
Problem presenter: John Peach
Faculty mentor: Minh Pham
Students: Hossein Aghakhani, Jingnan Fan, Alex Farrell, Ya-Ting Huang, Benjamin Levy, Het Mankad, Michael Minner

Water purification via Membrane Separation
Problem Presenter: Matthew Farthing
Faculty mentor: Lea Jenkins
Students: Fei Cao, Caleb Class, Tyson Loudon, Monica Nadal-Quiros, Star-Lena Quintana, Benjamin Ritz, Xiangming Zeng

Geographic and Racial Differences of Persons Living with HIV in the Southersn United States
Problem Presenter: Simone Gray
Faculty mentor: Howard Chang
Students: Isabel Chen, Christina Edholm, Rachel Grotheer, Tyler Massaro, Yiqiang Zhen

Microbes and Molecules: A Microscopic Analysis of Asthma
Problem Presenters: Agustin Calatroni, Russ Helms, Herman Mitchell
Faculty mentor: Sanvesh Srivastava
Students: Alexej Gossmann, Tamra Heberling, Nancy Hernandez Ceron, Yuanzhi Li, Anastasia Wilson, Hongjuan Zhou

Analysis of Self-Reported Health Outcomes Data
Problem Presenter: Mark Wolff
Faculty mentor: Kenny Lopiano
Students: Obeng Addai, Karianne Bergen, Shrabanti Chowdhuri, Fatena El-Masri, Xin Huang, Piaomu Liu

Chelyabinsk Meteorite Search via Image Segmentation

Hossein Aghakhani[1], Jingnan Fan[2], Alex Farrel[3], Ya-Ting Huang[4],
Benjamin Levy[5], Het Yagnesh Mankad[6], Michael F. Minner[7]

Faculty Mentors: John Peach[8], Minh Pham[9]

Abstract

On February 15, 2013 a meteor exploded in an airburst south of the city of Chelyabinsk, Russia and fragments of varying sizes scattered across the region. This project proposes a search algorithm for identifying sizable impact craters on the Earth's surface with the goal of locating unrecovered pieces of the Chelyabinsk meteor. We incorporate simulated debris fields of the event and tools from search theory in order to conduct a probabilistic search of the region. Our scanning sensor analyzes Google Earth™ images of the area and locates crater-like objects in the scenes via image segmentation by a level set method. By processing the images from before and after the event at each location, we are able to more accurately identify possible craters instead of simply identifying circular objects. This procedure reduces the number of false alarms that are detected in comparison to only using images from after the event. Our experimental results support previous estimates that most of the Chelyabinsk meteor evaporated during the airburst event and that it is unlikely a large, unrecovered fragment exists.

1 Introduction

On February 15, 2013 a meteor exploded in an airburst at about 23 km altitude and 40 km south of the city of Chelyabinsk, Russia. The explosion was responsible for numerous injuries and structural damages, though, thankfully, no deaths. The meteor is estimated to have possessed an entry mass in excess of $10,000$ tons. After the airburst, surviving fragments scattered across the region. One large piece in particular, with a mass of roughly 570kg, was excavated from the nearby Lake Chebarkul. Over 70% of the original mass of the meteor is estimated to have evaporated during the explosion, between 0.03% to 0.05% of the mass is estimated to have survived as meteorites, and the remaining mass was likely converted into dust.

Searching for objects is, in general, a difficult task, whether as part of a Search and Rescue operation or simply for target detection and identification. The mathematical theory of searching can vastly improve one's capability for finding an object of interest. In this project, we employ techniques of searching and image processing to scan an area of land near Lake Chebarkul for impact craters from the meteor. We utilize simulated debris data to develop a probabilistic model of the region in order to determine where these craters are most likely to be found. Relying upon image segmentation using a level set method, we develop a sensor to scan images of the region and locate crater-like objects in the scene. After fine-tuning the sensor with training data, we run the sensor on processed images from Google Earth™. The search pattern is dictated by a probabilistic model and the process terminates after all remaining regions possess a probability of containing a crater that is less than a specified threshold.

This work achieves three main results. The image segmentation approach that we employ is able to identify not only circular objects in an image, but also those objects which intersect the boundary of the picture. By processing images from before and after the airburst event, we are able to more accurately identify possible craters instead of simply identifying circular objects. This pre-processing method reduces the number of false alarms that are detected from the case where only images taken after the meteor event are considered. Our

[1]Mechanical and Aerospace Engineering Department, SUNY at Buffalo
[2]Center of Operations Research, Rutgers University
[3]Mathematics Department, Arizona State University
[4]Applied Mathematics and Statistics Department, Stony Brook University
[5]Mathematics Department, University of Tennessee
[6]Department of Mathematical Sciences, The University of Texas at Dallas
[7]Mathematics Department, Drexel University
[8]MIT Lincoln Laboratory
[9]SAMSI

experimental simulations support previous estimates that most of the Chelyabinsk meteor evaporated during the airburst event and decrease the likelihood that a significant fragment of the meteor, aside from the piece found in Lake Chebarkul, survived the airburst and impacted the region south west of Chelyabinsk.

1.1 Background on Searching

One begins a search to find a missing item by determining a set of locations where it is most likely to be found based on all information available. Next, one visits theses sites, in order from most likely to least likely, and inspects the local area until the object is found or one decides that the item is more likely at another location. Any additional, useful information obtained along the way is incorporated into the search. This procedure is repeated as one visits each site, sometimes even searching the same locations multiple times, until one finds the object or one gives up the search. This intuitive, common-sense approach to searching has a strong mathematical foundation. The report presented in [16] is an excellent introduction to this theory.

Adopting the notation of [16], we introduce the concepts of the *probability of containment*, the *probability of detection*, and the *probability of success*. The probability of containment, P_C, measures how likely the object is within a specified area. The probability of detection, P_D, measures how likely the sensor will detect an object if it is contained in the area being searched. Note that this is a *conditional probability*, i.e. it depends on another condition being true. In this case, the condition is that the missing object is in the containment region. The probability of success, P_S, is a measure of how successful, or effective, a search of an area will be; hence, it is the product of the probability of containment and the probability of detection, i.e. $P_S = P_C \times P_D$. Naturally, the central objective of any search is to maximize the probability of success. Although P_C and P_S can be adjusted after each area is searched to improve the likelihood of finding the object, our algorithm is unable to accommodate such changes due to the deterministic nature of our sensor.

Search theory is useful in a number of scenarios aside from meteorite detection. The general problem of finding a person or object in a large area is commonplace. For example, Air France flight 447 disappeared over the South Atlantic in June 2009. Exhaustive search efforts commenced in hopes of finding the wreckage but were unsuccessful. The authors of [23] devised a mathematically based search plan, incorporating the fundamentals of search theory, that quickly resulted in the discovery of the wreckage. Similar techniques were adopted to locate missing hydrogen bombs after the 1966 Palomares Incident [22]. More recently, Malaysian Airlines flight number 370 disappeared during a trip from Kuala Lumpur to Beijing and has yet to be found[27]. In such examples, search theory can be employed to better understand the problem, create hypotheses for debris locations and formulate and execute a search plan.

1.2 Literature Review

A great deal of work has been completed in researching the Chelyabinsk meteor as well as appropriate image processing and detection methods for finding the ejected fragments. We highlight a non-exhaustive sample of this ever-expanding body of work. For a comprehensive account of the airburst event and a detailed analysis of the meteor, see [20]. Alternatively, for a thorough investigation of the pre-impact orbit, consider [13]. The authors of [26] and [11] attempted to recreate the trajectory of the meteor. Both papers relied on available video and sensor data obtained during the entry of the object to estimate the relevant parameters of its approach and formulated three possible trajectories. Following this work, the authors of [18] then used the Planetary Entry Toolbox developed by [21] to simulate the debris fields that resulted from the estimated parameters of [11] and [26].

Adopting the relevant trajectory and debris fields from these articles, we develop a method to detect potential craters. In [15], researchers used Hough transformations to process and detect edges within an image; however, while applying their work in combination with various MATLAB functions and detection algorithms we find these methods perform poorly due to the diversity of the Earth's surface. Hence, the design of our sensor is based on the approach of the image segmentation by level set method as presented in [12].

1.3 Outline of the Paper

We begin by describing our problem and general approach in Section 2. This includes the assumptions that we make as well as details on the data we use and how we obtain it. Section 3 consists of in-depth details for how we address the problem. First the hypotheses and the corresponding probability map are constructed. We then provide information on the development of a sensor capable of detecting craters using satellite images. To improve the performance of the sensor we describe a method to process the images which also reduces the number of false alarms. With everything in place, we test the sensor on images of known craters, partial craters and non-craters, to assess its accuracy and overall performance. We finish the section by outlining the search procedure. In Section 4, we detail the main experiment and discuss the results. Finally, we state the conclusions of this work.

2 The Problem

2.1 Description

Although a fairly large piece of the meteor was found in Lake Chebarkul, most of the meteor is estimated to have evaporated. We consider the possibility that another large piece survived the airburst event but has not been found. Thus, we develop a crater detection method, determine an appropriate area to search, scan the area and analyze the results. Each part of this process requires careful consideration.

There are a number of uncertainties related to the meteor's trajectory, its fragmentation and the survivability of the object. The exact size of the meteor, its heading angle, the altitude and location when the object broke apart, forces acting on the object and the energy released when it exploded are all unknown. As a result, the precise trajectory and landing locations are not known, which complicates the task of determining which regions should even be searched. To make matters worse, there is no consensus on how much of the meteor survived the explosion. Furthermore, if any of the meteor survived other than what was recovered from Lake Chebarkul, there is no guarantee that the pieces would be large enough to create a detectable impact crater.

For the purposes of this project we assume that a large enough piece survived and produced an impact crater. We consider three scenarios in which the meteor fragmented into a large number of pieces and one in which the meteor remained largely intact post-airburst. These possibilities form the basis of our search area and probability map. However, even after narrowing down the search locations, we must still scan a 30 km $\times$ 80 km area of diverse terrain. A convenient approach to searching such a large region is to use satellite images captured by Google Earth™. We devise an automated process to guide this imaging service to the proper locations and store pictures of each site for later use. We apply a threshold on the probability map to further limit the number of locations to image Google Earth™. Once we have the pictures, we then need to scan them with a sensor to detect craters within each site.

Building the sensor is perhaps the most difficult aspect of this work. Although various researchers have worked on detecting craters on the Moon and Mars using satellite images, their methods do not directly apply to this project as the Earth's surface has considerably more features. We are unaware of other published works which attempt to search for meteorite craters on the Earth's surface using satellite images. Hence, we constructed a new method to accomplish this task. While we develop and test several different methods for crater detection, the underlying principle in each method is to search for circular objects in each image and determine whether or not the feature is a crater. Additionally, we apply processing procedures to the images in a way that will eliminate useless areas and accentuate crater-like features that can be detected. After designing the sensor, we assess its performance by testing it on various images that contain known craters and we record its success rate.

Once we determine the locations to search and scan the corresponding images with our sensor, the results are carefully analyzed. If potential craters have been found, we scrutinize the image for true detections, missed craters and false alarms. Once this process has been completed for the entire region, we multiply the calculated probability of detection with the probability of containment to determine the odds that a crater from the meteor does, or does not, exist in the region.

2.2 Assumptions

Since the precise trajectory of the meteor can only be estimated, we incorporate four possibilities when creating our probability map. Three of the trajectories were used by [18] to generate a simulated debris field under the assumption that significant fragmentation of the meteor occurred. Although this is the most likely scenario, we also consider a fourth trajectory determined by [19], which was created under the assumption that no significant fragmentation occurred. Regardless of the size of the meteor, we assume that any detectable impact resulted in a circular-shaped indentation in the Earth's surface that can be characterized by a steep gradient around the edges of the resulting crater. Based on this assumption, we attempt to detect the edges using several image processing and edge detection techniques via MATLAB.

2.3 Data

We begin the initial data collection process by compiling information on possible locations of debris from [18], which is based on the work by [26] and [11]. The data from [18] comes in the form of an image displaying simulated debris fields under the assumption that the meteor fragmented significantly. This image is read into MATLAB as an RGB file and locations of debris are marked. The locations are then translated into relative distances between debris, which is then used to create part of a probability map based on the clustering of points. Using the same image, we are also able to determine the longitude and latitude of the search area through careful measurement and segmentation of the displayed axis. The probability map is further developed using a fourth possible trajectory which was proposed by [19], as this is the most likely location that the remainder of the object would have reached assuming no significant fragmentation occurred.

We utilize the Google Earth™ API to generate images for a given latitude, longitude, altitude and range. Here the altitude is the distance from earth's surface in meters and the range is the distance in meters from the point specified by the longitude, latitude and altitude location, i.e. it is simply the zoom-in level. We also employ Python, shell script, ImageMagick and Apple Script for automating the process of fetching the images from Google Earth™. Passing a list of values (latitude, longitude, altitude and range) through a Python script to the Google Earth™ API script, we are able to quickly capture a large number of images. These images are then imported into the Crater Detection algorithm to search for possible craters. In order to reduce the probability of false detections, we also use images of the same locations prior to the meteor event. This allows us to compare the pre-event and post-event images of the same sites and perform additional processing to improve crater detection.

3 The Approach

Our approach to solving this problem involves seven steps:

1. Hypothesize possible meteor trajectories,
2. Develop a probability density map for our search,
3. Create and test a sensor to detect a meteorite crater,
4. Acquire images of each site via Google Earth™,
5. Process the images,
6. Design a search procedure to scan the area from regions of higher probability to less probability,
7. Analyze the results of the search.

3.1 Hypotheses, Probability Maps

Although the meteor plummeted to the earth in plain sight during the day, it was not detected by any scientific instrument prior to entering the atmosphere. As a result, its precise trajectory is unknown and thus cannot be used to create a probability map to guide our search algorithm. Furthermore, we have to consider several possible trajectories and the likelihood that the area surrounding each trajectory contains a meteorite. In

order to integrate these conditions into the construction of our probability map of the region, we turn to the works of [18] and [19].

In [18], the authors used the Planetary Entry Toolbox, developed by Elecnnor Deimos, to simulate several trajectories and the resulting debris fields from the meteor [21]. The Planetary Entry Toolbox is a collection of tools that was initially developed to simulate spacecraft reentry and has since been generalized for analysis and simulation of asteroids entering earth's atmosphere. Within the Planetary Entry Toolbox, the Endo-Atmospheric Simulator was paired with the DEBRIS tool to simulate such an entry as well as the resulting debris field post fragmentation. While Monte Carlo approaches are used in these tools to reduce uncertainties, all of the environmental inputs and parameter values must be supplied by the user. Since specific parameter values were not recorded, the authors adopted the work of [26] and [10] to simulate three independent entries. Although each simulation agreed on many parameters, they differed from one another in terms of velocity, altitude of fragmentation and heading angle. The resulting simulated debris fields are displayed in Figure 1.

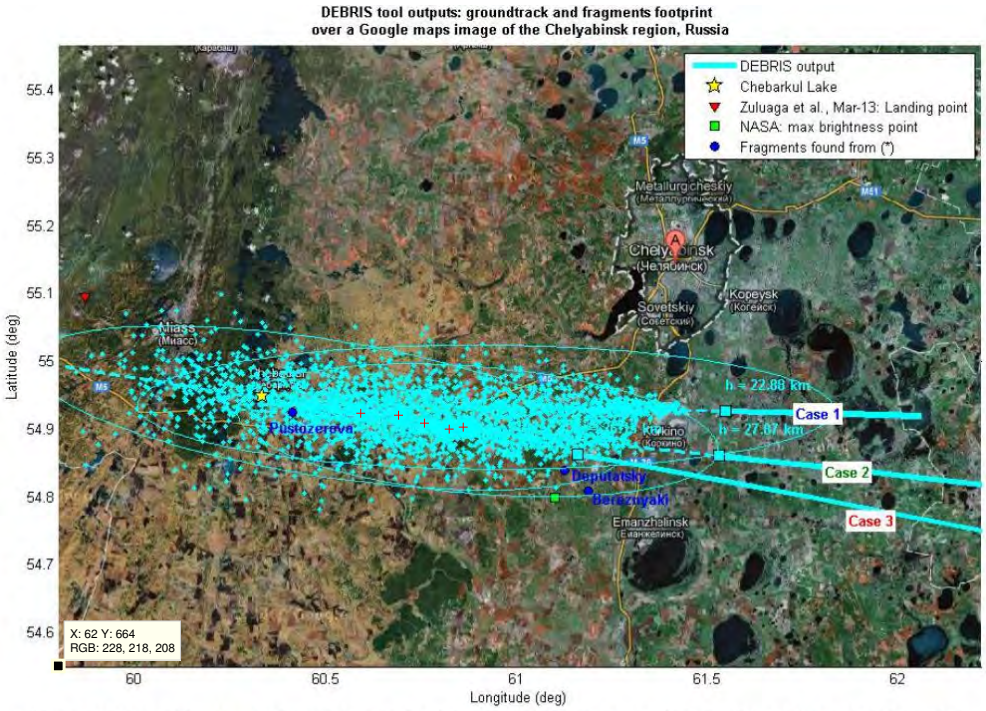


Figure 2 Groundtrack and fragment footprint over a Google Maps image of the Chelyabinsk region

Figure 1: Resulting debris fields from three independent simulations. Researchers from [26] and [11] considered different parameters to reconstruct three possible trajectories of the Chelyabinsk meteor. The authors of [18] then used their findings to simulated the meteor event using the Planetary Entry Toolbox from [21]. We use the image to create our probability map.

Incorporating the three trajectories, we create a probability map of the area based on where the debris fields are clustered. We begin this process by reading the image into MATLAB as an RGB file. The locations of the debris are marked by a 0 in the file, while all other entries contain a 1. Employing the *bwdist* command, we calculate the minimum Euclidean distance from each of the 0 entries to the closest nonzero entry. This effectively measures where the simulated debris fields are concentrated in relation to other debris locations. We intend to search areas that are most likely to contain debris; thus, we translate the information provided by the *bwdist* command into a probability map based on the relative distances as shown in Figure 2.

The sizes of the potential meteorites that we are searching for will impact our searching algorithm, as larger objects may travel a greater distance after the fragmentation. For example, Figure 1 illustrates how the largest known piece of the meteor, which was found in Lake Chebarkul, is located on the western side of the debris fields. In order to account for the possibility that a very large object continued past the debris field, we assimilate the estimation performed in [19]. This work utilized video footage taken during the meteor landing to estimate the angle, altitude and velocity of the trajectory and the forces acting on it in order

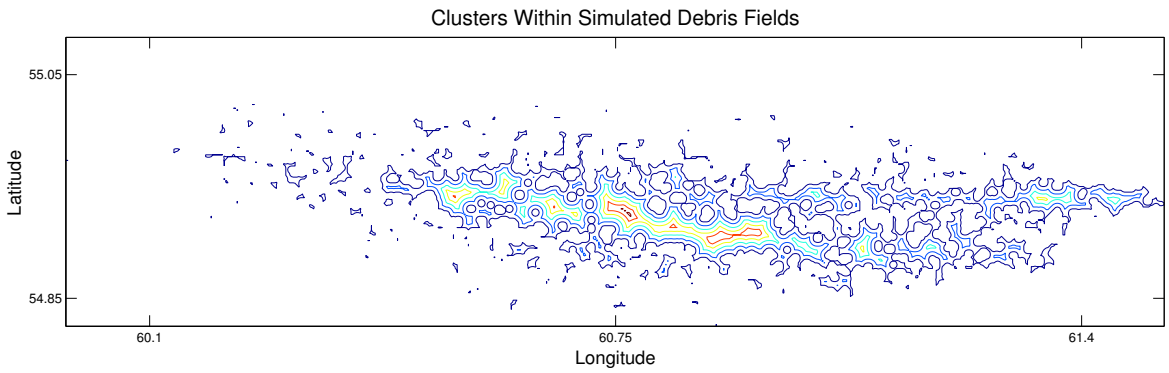


Figure 2: Contour Plot of Debris Clusters. The image from [18] is read into MATLAB and the locations of debris are marked in a matrix. The minimum distance from each debris location to the closest non-debris location is calculated and displayed as a contour plot. These distances are then used to create our probability map.

to determine a potential final resting point of the object. Considering the altitude of the landing site, the author estimated that the meteor could have made it as far as 55.0578°N latitude and 59.3197°E longitude, see Figure 3. Combining this location with an estimate of the standard deviation of the path, we create a normal distribution around this point and weight it the same amount as the other high probability areas in our map in order to generate the comprehensive probability map, which is illustrated in Figure 4. The hypothesized sites are listed in Table 1.

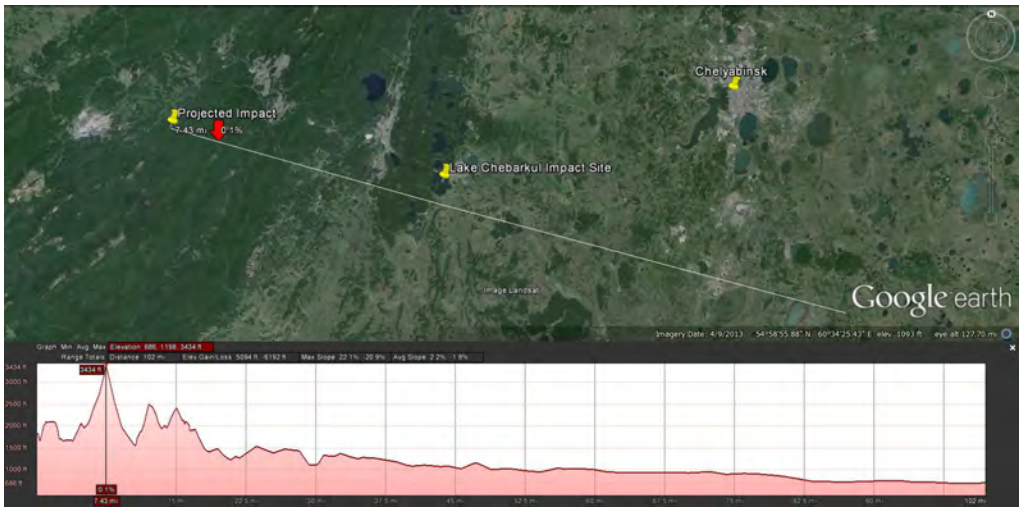


Figure 3: Estimated Flight Path and Landing Location of a Large Piece of the Meteor [6]. [19] The path displays the calculated trajectory of the meteor assuming a large piece of it remained intact after the explosion. (Google and the Google logo are registered trademarks of Google Inc., used with permission.)

3.2 Sensor

We consider several different approaches for designing an appropriate sensor to scan each image for craters. Unfortunately, two of these algorithms underperform in testing ultimately deciding to use an image segmentation by level set method. Our first approach for detecting circles, which we refer to as Identify Round Objects (IRO), requires the image be converted into black and white (binary). This process of transforming a

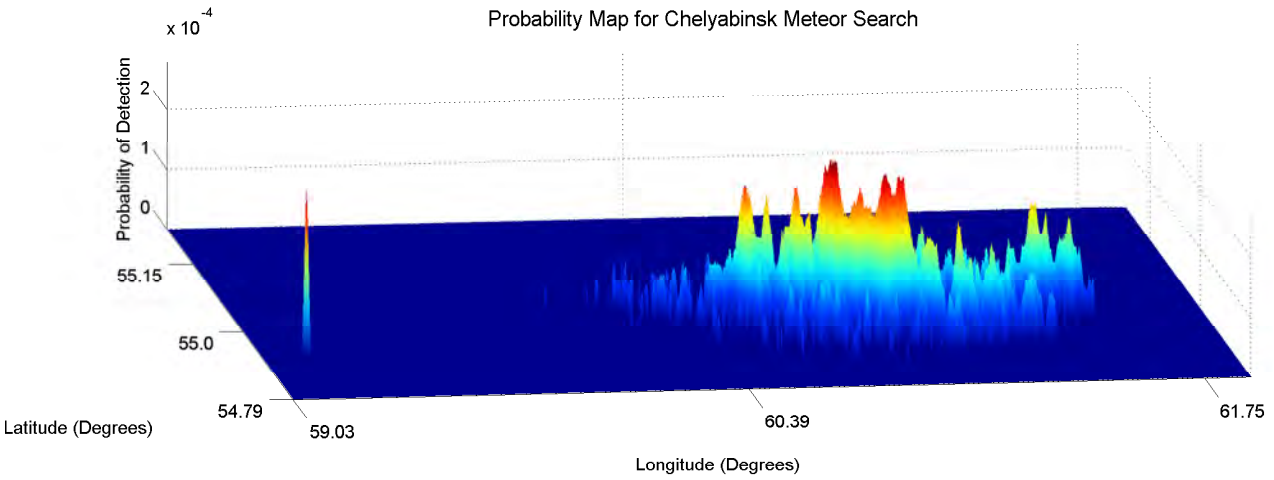


Figure 4: Probability Map of Potential Locations of Meteorites. The image from [18] is read into MATLAB and the locations of debris are stored in a matrix. The minimum distance from each debris location to the closest non-debris location is calculated. Since we wish to search areas that are most likely to contain a meteor, these distances are translated into a probability map based on how far each location is from areas without debris.

	Debris 1	Debris 2	Debris 3	Debris 4	Debris 5	Sizeable
Longitude	60.4507	60.5880	60.6865	60.8161	60.8653	59.1554
Latitude	54.9289	54.9203	52.9159	54.8957	54.9000	54.8797

Table 1: Longitude and latitude locations of the six highest probability hypothesis target sites in order from most likely to least likely. Using the image from [26], a probability map is created based on where the simulated debris fields are clustered. Five of the locations are situated within the simulated debris fields, which assumes the meteor fragmented significantly. The sixth location is based on [19] under the assumption that a very large piece of the meteor remained in tact post fragmentation.

picture with color to binary requires a threshold parameter of brightness to determine which pixels are sent to black. This parameter is key to obtain an image which retains essential details, as a poor threshold can result in a pure white or pure black picture. We apply additional image processing techniques to connect boundaries and fill holes in the binary image. IRO identifies how circular an object is by calculating the quantity $4\pi \times \text{Area}/\text{Perimeter}^2$ for each object. This will equal 1 if and only if the object is a circle, and it will be less than 1 otherwise. Hence, the closer this quantity is to 1, then the closer the object is to being a circle. Since this ratio uses the square of the perimeter, it will incorrectly identify several circular objects. For example, consider an object which is circular in nature but possesses a sinusoidal or saw-tooth perimeter. The object's perimeter is larger than the perimeter of a circle of comparable size; thus, the square of its perimeter will heavily impact the quantity calculated by IRO and it will not be identified as a circle. This approach requires excessive fine-tuning, which is not possible due to the scope of our project. Furthermore, even with additional pre-processing techniques, the IRO method was too coarse for our detection algorithm as it removed crucial information in order to produce results.

Next we employ a Circle Detection method using the Hough Transformation as detailed in [24] to search the images for craters. This approach combines the Generalized Hough Transform [15] with some basic image preprocessing (edge-detection via thresholding) to detect connected pixels in an image. Three points found on a connected edge are then used to calculate the parameters of a circle through the points. Although this method is successful at identifying craters on the moon, due to the uniformity of its surface, it is far less successful at detecting craters on the Earth. The primary source of errors, chiefly false positives, in our tests come from the diversity of Earth's surface, including trees, bodies of water and mountainous terrain. Often times the algorithm will detect discontinuities in an image, including the boundary of a crater, and fit distinct circles through points across these boundaries. Thus, instead of one large circle, it often produces numerous circles stitched together along both sides of the discontinuity. While this technique does provide control over several parameters, such as radii lengths, gradient threshold, etc., they also have to be finely tuned to each image in order to obtain reasonable results. Once again, due to the scope of this project and the number of images that needed to be scanned, we do not incorporate this method into our central algorithm.

The final method that we test for building our sensor is image segmentation by level set method. This method, introduced in [17], is a numerical method to track the topological changes of curves, surfaces and shapes and is used in many applications such as interface capturing in fluid dynamics, image segmentation in biological images and shape recognition. Different variations of this method have been developed, but the basic idea is to track the topological changes by solving an initial value partial differential equation (PDE), see (1). The level set function, ϕ, is a scalar signed distance function of position and time, and its value is assigned with respect to the curve position of the time. By signed distance function we mean that its value is equal to the distance from each point of the domain to the curve and its sign is positive if it is located outside of the curve and negative inside of the curve. Thus, given ϕ, we can find the position of the curve just by plotting the contour of $\phi = 0$ in the domain.

In right-hand side of equation (1), F is the normal velocity and $| \bigtriangledown \phi|$ is the magnitude of the gradient vector of ϕ. Thus, the change in ϕ, or the change in the closed curve, at each time is a function of the normal velocity and the gradient of ϕ. Figure 5 shows a signed distance function for an arbitrary shape [1].

$$\frac{\partial \phi}{\partial t} = F| \bigtriangledown \phi|, \qquad \phi(0, x, y) = \phi_0(x, y). \tag{1}$$

Depending on the problem, time t can be real or pseudo time. For example, typically in image processing when one wants to capture the image segments the time is pseudo time, which means that one starts with an initial guess and lets the algorithm run until it converges to the desired segments of the pictures. On the other hand, in transit fluid dynamics problems, time is the physical time and the level set represents the fluid interface at each time. To solve level set PDE, one has to have the initial level set ϕ_0. As discussed before, this initial level set must be a signed distance function. Given ϕ_0, which implicitly represents the position of the interface, we can update ϕ with a discretized time scheme, such an implicit Euler method. The level set ϕ only depends on the normal component of the velocity multiplied by the magnitude of the gradient of ϕ, because the normal velocity is the only component that could change the boundary position.

In many applications, like image segmentation, there is no physical velocity, so we have to use another parameter in place of the velocity in order to update the level set function. This is where several different

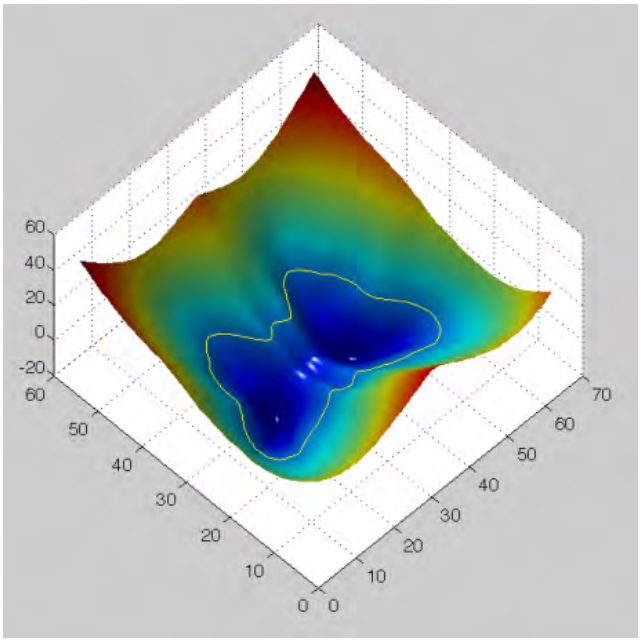


Figure 5: Signed Distance Function [1]: In this figure a signed distance function for an arbitrary shape is displayed. Notice the lower blue portion of the surface is negative, since it is inside the closed curve, while the upper portion of the surface is positive as it is outside the closed curve.

models can be invoked, for example one method is to set $F = \text{div}\left(\frac{\bigtriangledown\phi}{|\phi|}\right)$, which is the curvature of the curve. For more information about level set methods we refer the curious reader to [9].

The formulation of the level set method that we adopt in our sensor is different from the basic level set in that it is specifically developed for image segmentation [12]. This formulation enables it to detect objects whose boundaries are not necessarily defined by gradient and its derivation is based on the minimization of an energy functional. The result is presented below,

$$\begin{aligned} &\frac{\partial\phi}{\partial t} = \delta_\varepsilon(\phi) \quad \left[\mu \ \text{ div}\left(\frac{\bigtriangledown\phi}{|\bigtriangledown\phi|}\right) - \mathbf{n}\cdot\mathbf{u} - \lambda_1(u_0 - c_1)^2 + \lambda_2(u_0 - c_2)^2\right], \\ &\frac{\delta_\varepsilon(\phi)}{|\bigtriangledown\phi|}\frac{\partial\phi}{\partial\mathbf{n}} = 0, \end{aligned} \tag{2}$$

where $\mathbf{n}$ denotes the exterior normal to the boundary, c_1, c_2 are functions of ϕ and are Lagrange multipliers that minimize the energy functional, and $\mu \geq 0, \nu > 0, \lambda_1 > 0, \lambda_2 > 0$. The details of the derivation can be found in [12]. In this study, we used $\mu = 0, \lambda_1 = \lambda_2 = 1$. To implement this formulation of the level set method,we use the "Image segmentation - multi-scale energy-based level sets" code which is available in [14]. The code uses finite difference methods to solve the PDE and a multi-grid solver to more rapidly analyze the pictures.

3.3 Processing

Applying the sensor, we capture the edges of the image segments and the associated pixels, then we fit the best possible circle through these pixels. To ensure that the captured segment is very close to a circle, we measure the distance of each pixel to the center of the fitted circle and then calculate the standard deviation of the results normalized by the radius of the circle. This allows us to capture circles that intersect with the boundary of the image and discard those objects that are not circular nor a partial circle based on their curvatures. We also approximate the size of a probable crater with the estimated physical information available in [25].

Following these computations, the maximum diameter of the crater for this meteorite would be less than 500 meters, see Appendix. Considering that some parts of the meteor have been found in other places, we allow for the maximum diameter of 350 m for the crater. Moreover, based on the investigations of this meteor, we also estimate that the minimum size of crater that we could find is 50 m. This maximum and minimum threshold improve the ability of our sensor to more accurately find probable crater sites.

The "sensing" operation proceeds as follows:

1. Read the image into MATLAB,
2. Process the picture and find the image segments,
3. Fit the best possible circle through the points of the segment,
4. Use the threshold and the circle measure to distinguish between probable and impossible craters,
5. Store the radius and center of each circle,
6. Plot the selected circles on top of the image.

In order to enhance the sensor's crater detection capabilities, we develop an additional image processing technique, which we call "TimeDifferences." This approach analyzes a pair of pictures of each region, with one image from before the airburst and the other from after the event. The images are converted to gray-scale and then separately rescaled to possess the same gray level, which will account for differences such as seasonal disparities. From here we apply two processes, the first one is a gray to binary to binary difference map, where 'binary difference' entails taking the absolute value of the difference of the binary matrices. The result is a matrix that has white pixels where the two maps are very different and black pixels where the maps are roughly the same. The other method is a gray to gray difference to binary difference map, where the gray difference is a simple subtraction of the matrices which represent the images. While these both employ the same methods and result in a binary difference map, they yield surprisingly different results. We combine the two maps by taking the maximum of the entries in the two matrices (if either entry is a white pixel, the output is a white pixel) in order to create a binary mask. This mask is sent to a partial gray mask by changing the black to gray. Next we apply a Gaussian smoothing operation to the gray mask to obtain our final mask. This final mask is applied to our original picture from current time, and the resulting image is sent to our sensor, see Figure 6. The application of the final mask reduces the probability of crater detection in the areas which did not change over time. This pre-processing routine reduces the occurrence of false positives, see Section 4. Figure 7 illustrates the input and output of the image pre-processing on an example image, displaying the pre-event photo, the post-event photo, and the resulting shaded map.

3.4 Testing the Sensor

In order to test the sensor, we gather known crater site coordinate locations from around the world as provided in [5] and [4]. Here, we employ Google Maps™ to visit these coordinates and visually inspect them for craters. Some of the crater sites are now lakes, however, they still serve as reasonable tests for crater detection. Furthermore, several locations contain partial craters, others lack craters, and still others posses varying environments. Due to time constraints and the paucity of crater sites on Earth, we test the sensor on 45 of these images. First, we scan 15 sites at random and run the sensor on each image. We visually inspect the results to determine how many craters are correctly identified, how many are missed and the number of false alarms. Based on this information, we adjust the maximum and minimum diameters for acceptable crater sizes in order to improve the probability of detection. Next, we apply the sensor to the remaining 30 images and compare the results, as displayed in Table 2. Here we see that the probability of detection, P_D, has improved, but the false alarm rate has also increased. Examples of the sensor scanning test images are illustrated in Figures 8 and 9. Unfortunately, we are not able to obtain pre-impact and post-impact photos of these craters and thus can not perform the "TimeDifferences" processing on them. However, the results demonstrate the capability of the sensor to detect potential craters in an image.

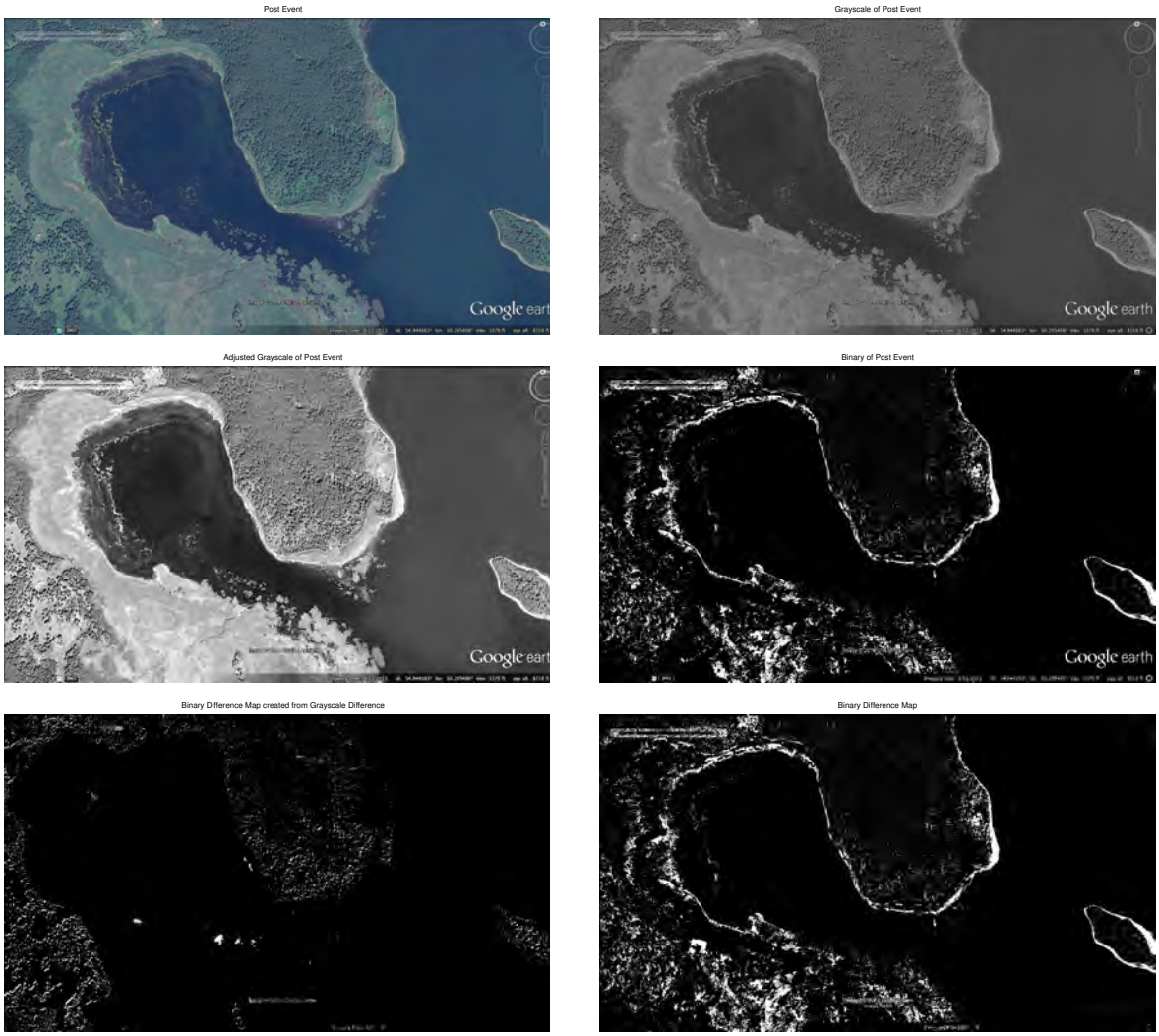


Figure 6: Photos Taken From Steps Within Pre-Processing Algorithm [8]. The photo is read in (Past Event), turned to grayscale (Grayscale), adjusted to have proper gray levels overall (Adjusted Grayscale), and then converted to black and white (Binary) in order to begin edge detection. The difference in the gray photo matrices is sent to binary (Grayscale $\rightarrow$ Binary Difference) and combined with the binary difference (Binary Difference Map) to create the binary mask. (DigitalGlobe. Google and the Google logo are registered trademarks of Google Inc., used with permission.)

	True Detections	Missed Targets	False Alarms	P_D	False Alarm Rate
Training Set - 15 Images	10	5	8	2/3	8/15
Testing Set - 30 Images	30	6	24	5/6	4/5

Table 2: Summary of results when training and testing the sensor. For each set of images, the associated number of correctly detected craters, missed craters and false alarms are displayed. These values are then used to calculated the probability of detection, P_D, and false alarm rate, which is the number of false alarms per the number of images.

Figure 7: Example of TimeDifferences Image [8]. The current photo (Middle) is compared with a photo taken before the meteor hit (Top), and gives us a new shaded colormap (Bottom) with regions of no change shaded to be darker than regions that have changed. (DigitalGlobe. Google and the Google logo are registered trademarks of Google Inc., used with permission.)

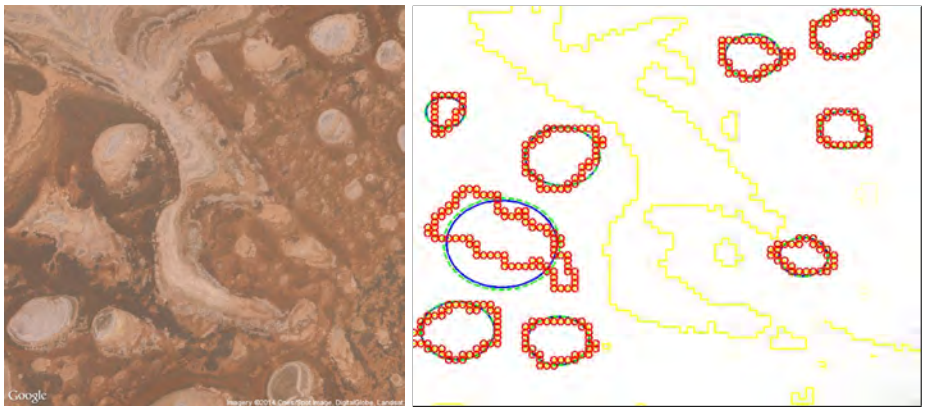


Figure 8: The Result of Analysis with the Sensor: The left image is from Google Maps™ , and the right image results from applying the sensor. The yellow parts are the boundaries of the image segments while the red lines are the best possible circle fitted through yellow lines that satisfy the defined circle measure and diameter threshold. Notice that the sensor successfully captures the possible craters in the image. (DigitalGlobe. Google and the Google logo are registered trademarks of Google Inc., used with permission.)

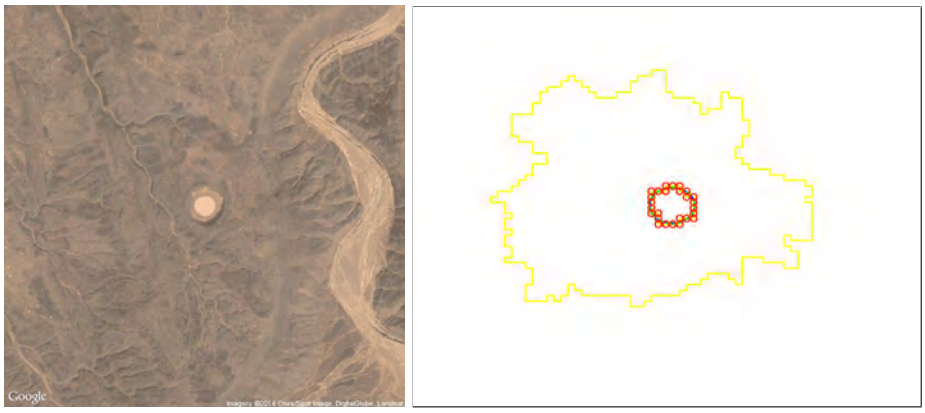


Figure 9: The Result of Analysis with the Sensor: The left image is from Google Maps™ , and the right image results from applying the sensor. The only possible crater is detected successfully. (DigitalGlobe. Google and the Google logo are registered trademarks of Google Inc., used with permission.)

3.5 Search Procedure

Analyzing the probability distribution for the region in question, we first discretize the probability map into cells with dimensions comparable to 1 km $\times$1 km in Google Earth™ . These dimensions were determined by our estimation of the crater size and computational capacity. Next, we calculate the total probability contained in each cell and re-order the cells from highest to lowest values. The longitude and latitude associated with the center of each cell are sent to Google Earth™ in order to obtain pre-event and post-event images of these sites from a fixed altitude of 1780 meters. This approach will allow us to evaluate the efficacy of the "TimeDifference" method as we can apply the sensor to the present day images as well as the images that result from the "TimeDifferences" procedure. Next, the images are read into MATLAB for processing. This approach enables us to scan areas of the region which are more likely to contain craters first. Since many cells possess a negligible probability, we set a threshold to zero out these regions and scan only the cells which account for the greatest 90% of the total probability distribution. For simplicity and reduced computational complexity, we do not apply a sliding window to scan overlapping regions around each area, which would enable the sensor to more accurately detect craters located on the boundaries of the original image. Instead, we move from cell to cell according to the probability distribution and never rescan the same cell due to the deterministic nature of our sensor, i.e. the sensor will never return a different output for the same images. Thus, we do not adjust the probability density function after searching a cell. This approach speeds up the procedure considerably.

4 Computational Experiments

The experiment applies the sensor described in Section 3.2 to a collection of pre-event and post-event images of locations specified by our probability map. The output of this algorithm is a set of potential crater sites. Our test data was obtained exactly as described in Section 2.3. Although these are not optimal images since the imaging services occasionally obtain blurred pictures, are unable to image a site pre-impact or suffer other degradations, they are the only available data to search the Chelyabinsk region for craters. The probability density map, see Figure 4 reduces the search area to only require 384 photos (768 with TimeDifferences) from Google Earth™.

The "TimeDifferences" processing method sends the white from our difference mask to 0.2 gray (on the scale from 0 to 1) and creates an 85×85 Gaussian matrix with mean 1 and standard deviation 20, as measured in pixels, in order to smooth out the partial gray mask. We normalize this mask by its maximum value and set the crater radius threshold to be between 50 and 350 meters. The first approach, which only uses present photos from Google Earth™ (Present Approach), returned 52 images that contained craters with a total of 71 craters across all images. Simultaneously, we run our second algorithm, also using Google Earth™ (Past Approach), which compares data from pre-event and post-event images via the "TimeDifferences" processing method. This approach returned 67 images with 98 craters detected in total. Note that for cases where the data provided by Google Earth™ for pre-event images was insufficient, we resort to the Present Approach. Any pictures for which the past data was inadequate or too degraded, were removed from the visual search, which further reduced the number of images to inspect down to 28 pictures with 34 craters in total.

We visually inspect these pictures for craters, see Figure 10 for an example of the output. The results, as displayed in Table 3, demonstrate an appreciable improvement when using past and current data. Analyzing the results of the Past Approach, we find 7 true detections, 4 missed targets and 27 false alarms; hence, we attain a probability of detection of $P_D = 7/11 \approx 64\%$. We conclude that there approximately is a 57% chance (90% chance the crater is in the search area multiplied by the 64% probability of detecting of a crater in a searched cell) that the crater did not yield any sizable meteor fragments, other than the Lake Chebarkul piece, that impacted the region displayed in Figure 4.

5 Summary and Future Work

Image processing via TimeDifferences can improve searching efficiency in areas where picture data at times before and after an event are available. This approach has the potential to reduce computational costs and the number of false positives identified in images where significant changes occur over time. The image

Figure 10: Example Detection with Sensor [7]: Our pre-event image (Top) has a full lake and field, while the post-event photo (Middle) shows a dry lake and no field. The differences assist the sensor in detecting both objects. (DigitalGlobe. Google and the Google logo are registered trademarks of Google Inc., used with permission.)

	True Detections	Missed Targets	False Alarms	P_D	False Alarm Rate
Present Approach - 52 Images	0	1	71	0	71/52
Past Approach - 28 Images	7	4	27	7/11	27/28

Table 3: Summary of experimental results. For each set of images, the associated number of correctly detected craters, missed craters and false alarms are displayed. These values are then used to calculated the probability of detection, P_D, and false alarm rate, which is the number of false alarms per the number of images.

segmentation by level set method performs well on the Earth's surface whereas other methods falter. Our results strengthen previous estimates that most of the Chelyabinsk meteor evaporated during the airburst event and that it is unlikely any other significant pieces impacted the region detailed in Figure 4.

A limitation of our method is the inability to distinguish between a crater and another circular object in certain situations. This is partially addressed by requiring a bound on the crater size so detected circles that are too big are discarded and by thresholding on the curvature of the object so irregular objects are ignored. An additional approach to further distinguish craters and other circular objects on the Earth's surface is to incorporate colormap discrimination to eliminate non-crater objects such as bodies of water and circular forests, however, one must beware of the strong diversity in the coloring of these bodies. Lastly, a database composed of higher-resolution images taken at regular, and more frequent, intervals will not only allow the sensor to more accurately detect and classify circular objects in each site but also extend the areas we can search to overlapping regions and find cropped impact craters.

Appendix

We can estimate the crater diameter via the following equation [2][3]:

$$D = 0.07 C_f \left(\frac{g_e}{g}\right)^{\frac{1}{6}} \left(W \frac{\rho_a}{\rho}\right)^{\frac{1}{3.4}}. \tag{3}$$

Here, we utilize the following quantities: C_f: Crater Collapse Factor (this is equal to 1.3 for craters > 4 km on Earth), g_e: Gravitational Acceleration at the surface of Earth, g: Acceleration at the surface of the body on which the crater is formed, W: Kinetic Energy of the impacting body (in kilotons TNT equivalent), ρ_a: Density of the impactor (ranging from 1.8 g/cm^3 for a comet to 7.3 g/cm^3 for an iron meteorite), ρ: Density of the target rock. For the Chelyabinsk meteor, we have the following information: Mass = 11,000 tons, v= 18.6 m/s, ρ_a=3600 kg/m^3.

Given this data, the diameter of the crater is 0.6 km. We verify this equation with other equations available in the literature. In order to determine an appropriate image size to find the impact location, we plot the result of this equation over a range of meteor diameters. Figure 11 shows the crater diameter obtained from the above equation as a function of the meteor's diameter and mass. Figure 12 highlights an approximately linear relation between the range of the meteor diameter and the crater diameter.

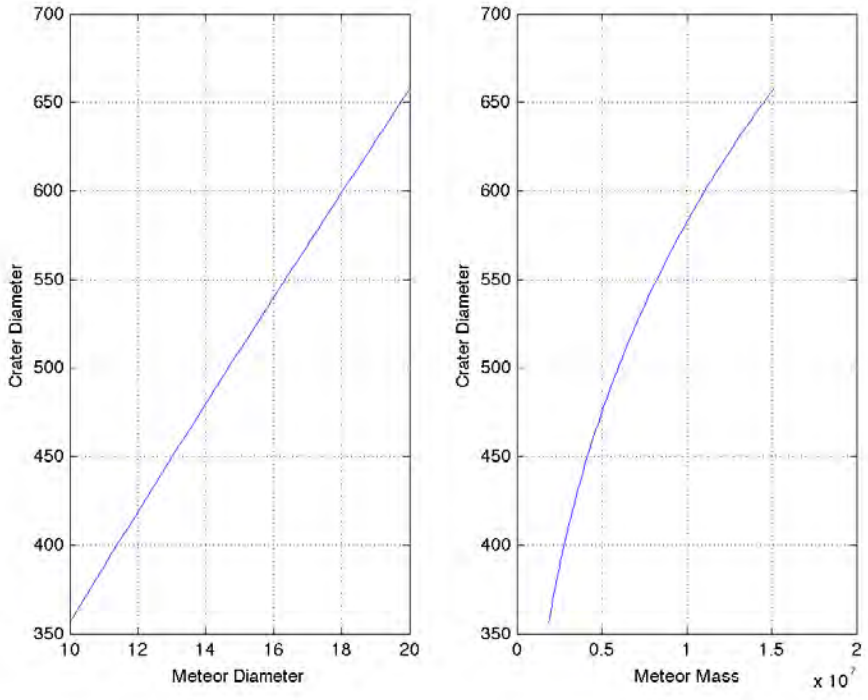


Figure 11: The Diameter of the Crater: The left figure shows the diameter of the crater as a function of meteor crater, and the right figure shows the crater diameter as a function of meteor mass.

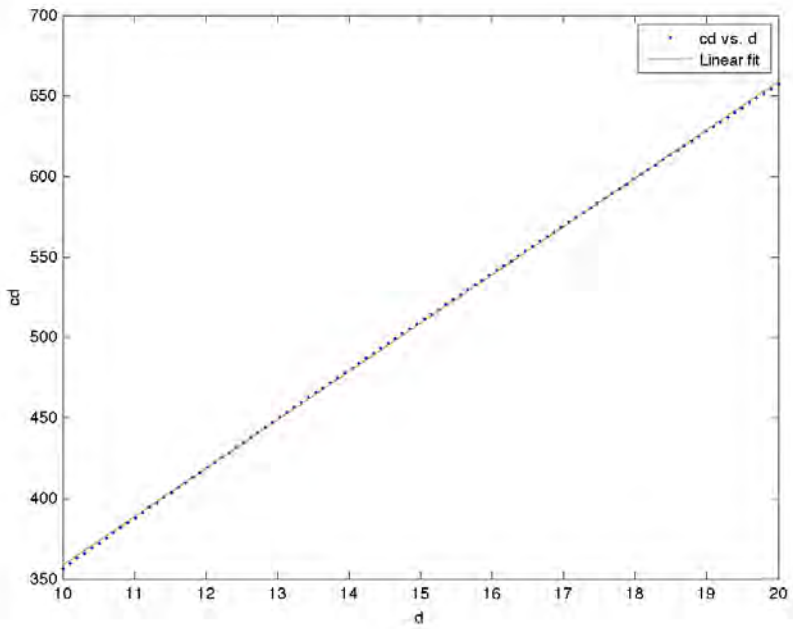


Figure 12: The Linear Fit Result: This figure displays a linear relationship between the crater diameter and meteor diameter.

References

[1] http://snikolov.weebly.com/uploads/3/9/0/4/3904101/945746.png?460l.

[2] http://palaeo.gly.bris.ac.uk/Communication/Brana/equation.html.

[3] http://www.lpl.arizona.edu/tekton/crater_p.html.

[4] Geoscience News and Information. http://geology.com/meteor-impact-craters.shtml, 2014.

[5] Scaredy cat films. http://impact.scaredycatfilms.com/, 2014.

[6] Google Earth 7.1.2.2041. Field and lake detection. lat 55.057825 lon 59.319739, Google Earth, September 4, 2013, July 22, 2014.

[7] Google Earth 7.1.2.2041. Field and lake detection. Google Earth, December 31,2013, July 22, 2014.

[8] Google Earth 7.1.2.2041. Scanned region near lake Chebarkul. lon 54.944083, lat 60.295408, Google Earth, August 12, 2013, Viewed July 22, 2014.

[9] Elsa Angelini, Yinpeng Jin, and Andrew Laine. State of the art of level set methods in segmentation and registration of medical imaging modalities, 2001.

[10] J. Borovicka, P. Spurny, P. Brown, P. Wiegert, P. Kalenda, D. Clark, and L. Shrbeny. The trajectory, structure and origin of the Chelyabinsk asteroidal impactor. *Nature*, 2013.

[11] J. Borovicka, P. Spurny, and L Shrebeny. Trajectory and orbit of the chelyabinsk superbolide. *Central Bureau for Astronomical Telegrams*, 3423, 2013.

[12] T.F. Chan and L.A Vese. Active contours without edges. *Image Processing, IEEE Transactions on*, 10(2):266–277, Feb 2001.

[13] C. de la Fuente Marcos and R. de la Fuente Marcos. Reconstructing the Chelyabinsk event: pre-impact orbital evolution. *MNRAS*, 443:L39–L43, 2014.

[14] T. Dima. http://www.mathworks.com/matlabcentral/fileexchange/31975-image-segmentation\
\-multiscale-energy-based-level-sets.

[15] R. O. Duda and P. E. Hart. Use of the Hough Transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, January 1972.

[16] U.S. Coast Guard Office of Search Rescue. *The Theory of Search: A Simplified Explanation.* Soza & Company, Ltd., 1996.

[17] S. Osher and J. A. Sethian. Fronts Propagating with Curvature-dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations. *J. Comput. Phys.*, 79(1):12–49, November 1988.

[18] C. Parigini, J. L. Cano, and R. Haya-Ramos. Preliminary estimation of the footprint and survivability of the Chelyabinsk Meteor fragments. *ArXiv e-prints*, 2013.

[19] J. Peach, 2014. Pers Comm.

[20] O. P. Popova, P. Jenniskens, et al. Chelyabinsk airburst, damage assessment, meteorite recovery, and characterization. *Science*, 342(6162):1069–1073, 2013.

[21] R.H. Ramos and L.F. Penin. Lanetary Entry Toolbox: A SW suite for Planetary Entry Analysis and GNC Design. October 2006.

[22] D. Stiles. A Fusion Bomb over Andalucía: U.S. Information Policy and the 1966 Palomares Incident. *Journal of Cold War Studies*, 8(1):49–67, 2006.

[23] L. Stone, C. Keller, T. Kratzke, and J. Strumpfer. Search Analysis for the Location of the AF447 Underwater Wreckage. `http://www.bea.aero/fr/enquetes/vol.af.447/metron.search.analysis.pdf`, 2011. Report to Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile.

[24] A. Upadhyay. Crater boundary detection and size frequency distribution (SFD) using MATLAB®, 2011. Indian Institute of Remote Sensing, Indian Space Research Organization.

[25] D. Yeomans and P. Chodas. Additional Details on the Large Fireball Event over Russia on Feb. 15, 2013. `http://neo.jpl.nasa.gov/news/fireball_130301.html`, 2013.

[26] J. I. Zuluaga and I. Ferrin. A preliminary reconstruction of the orbit of the Chelyabinsk Meteoroid. *ArXiv e-prints*, 2013.

[27] J. Zweck. How did Inmarsat deduce possible flight paths for MH370? *SIAM News*, 47(4), 2014.

Simulation-Based Optimization of Membrane Performance

F. Cao[1], C. Class[2], T. Loudon[3], M. Nadal-Quirós[4], S. Quintana[5], B. Ritz[6], and X. Zeng [7]

[1]Pennsylvania State University
[2]Massachusetts Institute of Technology
[3]Colorado School of Mines
[4]University of Puerto Rico
[5]Temple University
[6]Clarkson University
[7]North Carolina State University

July 22, 2014

Abstract

Often times, filtration is used to remove a particular contaminant; however, using a separation process in conjunction with filtration can provide an alternative solution that will allow for the retrieval of valuable components. The separated material can go on to be used for other purposes, such as polymer processing and pharmaceutical manufacturing. A simulation-based optimization was built to understand how to improve membrane performance for filtration and separation processes. Given equilibrium concentrations and isotherm parameters, a 1-dimensional transport model approximated the amount of adsorbate in the adsorbent at equilibrium by using single- and multi-component isotherm models. A global optimization algorithm that combined a simulated annealing method with a descent method was developed to optimize the nonlinear least-squares approach on which the inverse problem is based. In addition, two distinct applications were examined: water purification, where the volume of purified liquid is maximized, and protein recovery for pharmacology, where the volume of biological product is maximized efficiently.

1 Introduction

Filtration and separation processes play an important role in a variety of industrial applications, ranging from pharmaceutical manufacturing to polymer processing to water purification. Often the application requires removal of a contaminant (e.g., water purification), but separation processes are also developed to retrieve components of solutions needed to develop other compounds or materials (e.g., obtaining biological proteins from animal sera).

The dynamics of separation processes still need further research due to the complex interactions between solutes and membranes [4, 13]. Due to improvements in computing efficiency, numerical models provide a way of understanding these separation processes. Isotherms are important aspects of a simulation process, because they describe the equilibrium of adsorption of a material on a surface at constant temperature. The separation processes are highly affected by the isotherm properties; however, the explicit expression of isotherms can only be extracted from limited experimental data. As a result, there are many parameterizations during the numerical simulation process, which would affect the separation process. In fact, there are also many other factors that would influence the efficiency of a separation process, such as injection velocity, membrane porosity, and filtering time. Some of these factors can be adjusted according to the objectives of different separation applications. Generally, multi-objective optimization methods can be used to find the best values of parameters corresponding to different applications, based on the numerical models and choice of isotherm expressions.

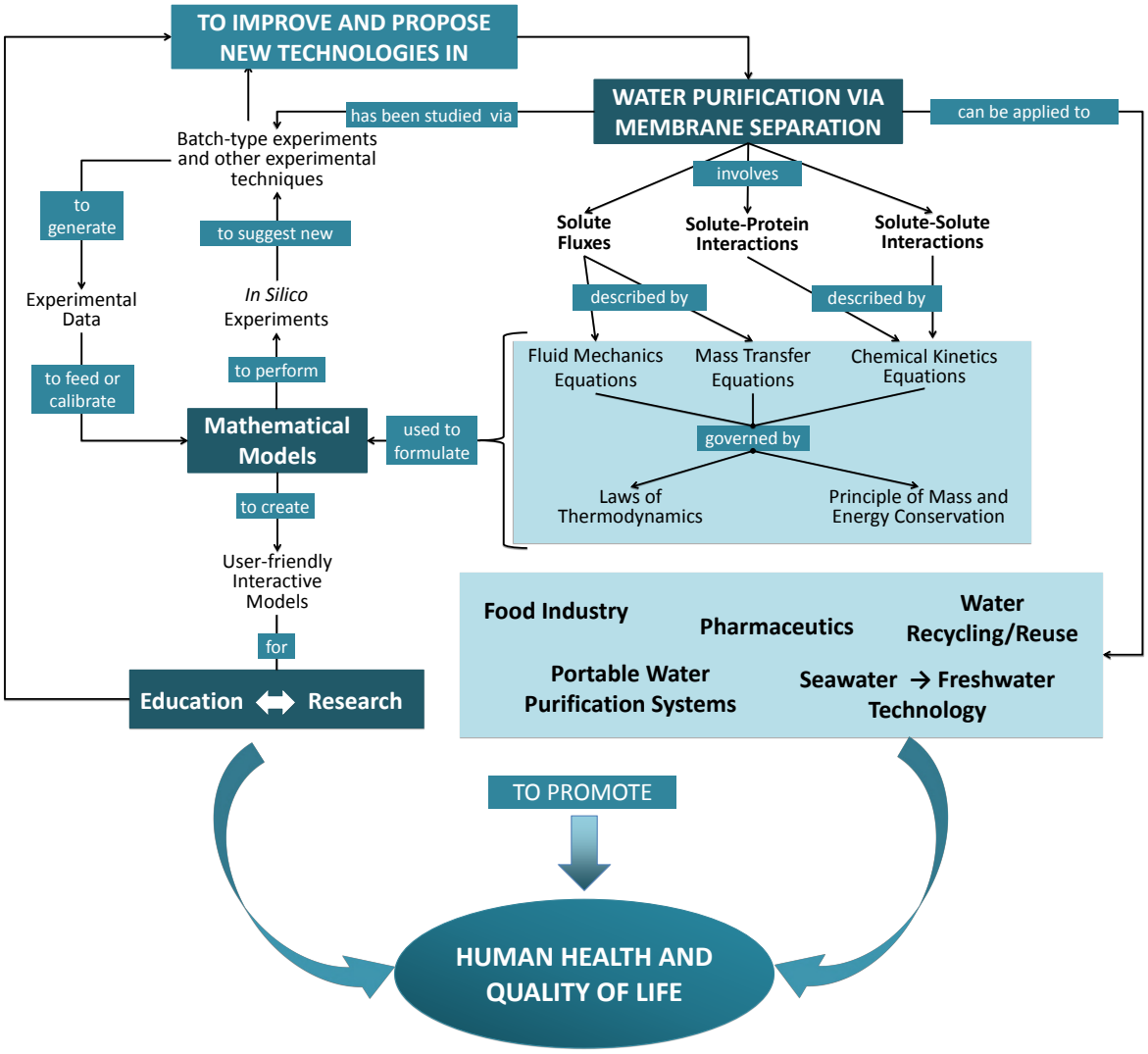


Figure 1: Concept map.

There are twelve important characteristics of water purifiers outlined by the US Army Center for Health Promotion and Preventative Medicine's (USACHPPM) Water Supply Management Program for war fighters [2]. Four of the listed characteristics are included in the optimization routine: reduce pathogens, purify quickly, remove taste/odor, and resist clogging.

In the mathematical model of the membrane, the continuity equation is used, which includes terms for convection, axial diffusion, and adsorption, as in [12]. Multiple isotherms are considered, which are listed in tables 4 and 5). Porosity of the membrane was treated as if it was in quasi-equilibrium as in [9]. In other words, for the discretized model, the porosity was treated as a constant for each time step. Then, at the end of each time step, porosity was updated with the constitutive relationship given in [9].

The study was based on the work by [13], in which Tarafder defines several objective functions that are modified for this study. In addition, extensions to multi-component isotherms are incorporated to better understand the connections between the selected design parameters and the overall performance of the membrane.

This paper provides the 1-dimensional mathematical model of the system (section 2) that was the basis for this study. Afterwards, improvements and tools used for optimization are included (section 3). Examined applications are included in section 4, along with particular constraints in section 5, followed by numerical resul
for fi

2

T
wher
targe
coul
phas
purp
mod

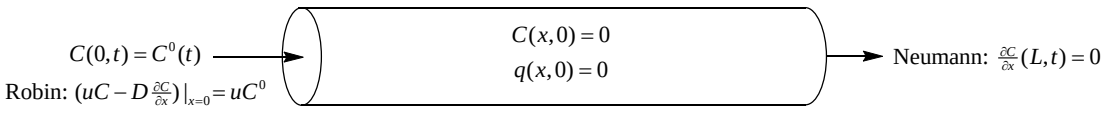


Figure 2: 1D flow with Initial and Boundary conditions

2.1 One component model

Before looking at the 2-component model, a 1-component model will be studied briefly in order to verify results. The fluid will be modeled as moving from left to right (see Figure 2). At the initial time, there is no solid phase mass on the membrane, $q(x,0) = 0$, nor any solute material anywhere in the fluid such that $C(x,0) = 0$. For the left boundary, the concentration of solute material is held as a constant, $C_(0,t) = C^0(t)$. For the right boundary, Neumann conditions are used $\frac{\partial C}{\partial x}(L,t) = 0$. For the 2-component model, the left boundary will change.

The method of lines can be used to find a solution to the 1-component mass conservation system:

$$\theta\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} - D\frac{\partial^2 C}{\partial x^2} = k_m(\theta - 1)\left[q^{eq}(C) - q\right] \tag{1}$$

$$\frac{\partial q}{\partial t} = k_m\left[q^{eq}(C) - q\right], \tag{2}$$

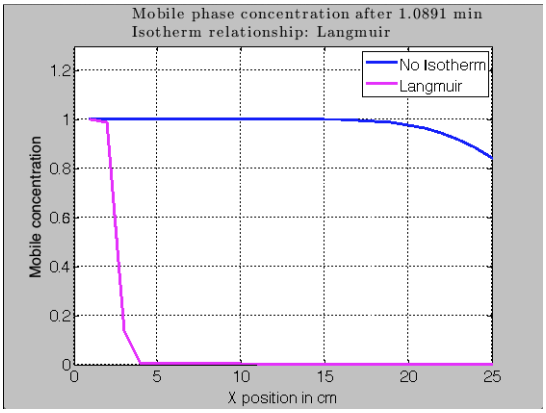


Figure 3: 1 Component- 1 minute

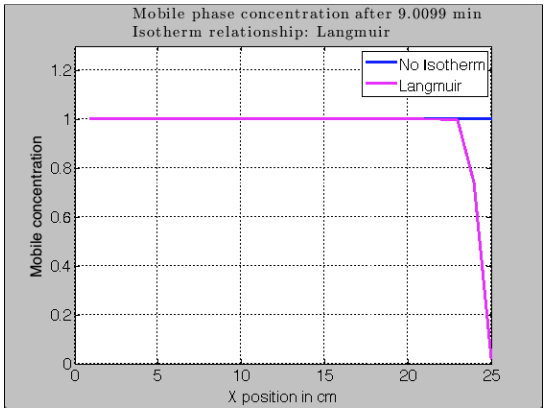


Figure 4: 1 Component- 10 minute

where

θ = porosity (volume fraction of the mobile phase) [-]
C =mobile phase concentration [mg/cm^3]
q = immobile (solid) phase mass fraction [mg/cm^3]
u = velocity of the mobile phase [cm/min]
D = effective mobile phase diffusion/dispersion coefficient $[\frac{\text{cm}^2}{\text{min}}]$
k_m = rate coefficient for mass transfer [min^{-1}]
q^{eq} = isotherm (solid phase equilibrium)

In order to discretize the advection term in Equation 1, an upwind scheme was used, while a centered difference scheme was used to discretize the diffusion term(here we assume that $u \leq 0$) (Table 1).

Table 1: Spatial Discretizations

$\frac{\partial c_i}{\partial x} = \frac{c_i - c_{i-1}}{\Delta x}$
$\frac{\partial^2 c_i}{\partial x^2} = \frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2}$

After setting parameter values to those listed in Table 2, choosing the Langmuir isotherm from Table 4 and isotherm parameters from Figure 13, a solution to the system was computed (Figures 3 and 4). Note: For the 1-component model, θ is a scalar.

Table 2: Parameter values for 1-component model

Parameter	Value
u	30 $\frac{\text{cm}}{\text{min}}$
D	0.005 $\frac{\text{cm}^2}{\text{min}}$
k_m	1000 m^{-1}
θ	0.74
L	25 cm
Δx	1 cm

2.2 Two component model

Now, the one-component model can be extended to two components in order to evaluate what happens when multiple components are interacting with one another.

$$\theta\frac{\partial C_1}{\partial t} + u\frac{\partial C_1}{\partial x} - D_1\frac{\partial^2 C_1}{\partial x^2} = k_{m,1}(\theta - 1)\left[q_1^{eq}(C_1, C_2) - q_1\right] \tag{3}$$

$$\frac{\partial q_1}{\partial t} = k_{m,1}\left[q_1^{eq}(C_1, C_2) - q_1\right] \tag{4}$$

$$\theta\frac{\partial C_2}{\partial t} + u\frac{\partial C_2}{\partial x} - D_2\frac{\partial^2 C_2}{\partial x^2} = k_{m,2}(\theta - 1)\left[q_2^{eq}(C_1, C_2) - q_2\right] \tag{5}$$

$$\frac{\partial q_2}{\partial t} = k_{m,2}\left[q_2^{eq}(C_1, C_2) - q_2\right] \tag{6}$$

with

C_i =mobile phase concentration [mg/cm^3] of component i
q_i = immobile (solid) phase mass fraction [mg/cm^3] of component i
D_i = effective mobile phase diffusion/dispersion coefficient for component i [$\frac{\text{cm}^2}{\text{min}}$]
$k_{m,i}$ = rate coefficient for mass transfer of component i [min^{-1}]
q_i^{eq} = isotherm (solid phase equilibrium) for component i

For the two-component model, we can choose either the Dirichlet or Robin condition as the boundary condition of the left-hand side. Here, we choose Robin under the main consideration that this type better matches the physical conditions in the actual processes. (Figure 2). The sum of the advective and diffusive fluxes at the boundary will be equal to the velocity of the mobile phase multiplied by the initial concentration of the respective component in this phase. The left-hand side of equation 3 can be discretized when the advective and diffusive fluxes are combined into a total flux, $f_{i-\frac{1}{2}}$:

$$\frac{\partial}{\partial t}\left[\theta C_i\right] + u\frac{\partial C_i}{\partial x} - D\frac{\partial^2 C_i}{\partial x} \tag{7}$$

$$u\frac{\partial C_i}{\partial x} - D\frac{\partial^2 C_i}{\partial x^2} = \frac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{\Delta x} \tag{8}$$

$$\implies \frac{\partial}{\partial t}\left[\theta C_i\right] + \frac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{\Delta x} \tag{9}$$

For $f_{i+\frac{1}{2}}$,

$$f_{i+\frac{1}{2}} = uC_i - D\left(\frac{C_{i+1} - C_i}{\Delta x}\right) \tag{10}$$

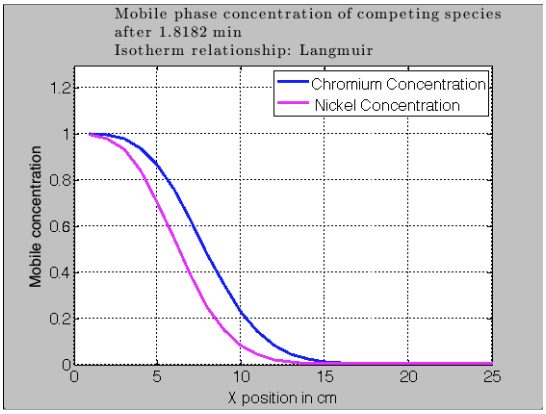


Figure 5: 2 Component- 1 minute

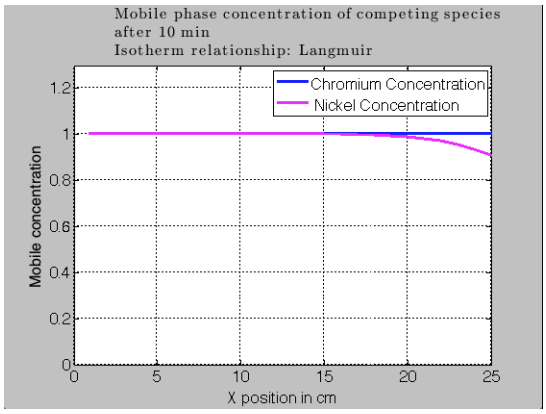


Figure 6: 2 Component- 10 minute

For $f_{i-\frac{1}{2}}$,

$$f_{i-\frac{1}{2}} = uC_{i-1} - D\left(\frac{C_i - C_{i-1}}{\Delta x}\right) \tag{11}$$

But when the Robin boundary condition is left as:

$$(uC - D\frac{\partial C_i}{\partial t})|_{x=0} = uC^0 \tag{12}$$

Then the corresponding flux $f_{i-\frac{1}{2}}$ for $i = 1$ would be changed to:

$$f_{i-\frac{1}{2}} = uC^0, \text{where } C^0 = C^0(t) \tag{13}$$

The assumed values of u and D are constant, while θ will be considered a function of time. Initial conditions are also assumed (Table 3).

Table 3: Initial and Boundary Conditions for each component i

$$\begin{aligned} C_i(x,0) &= 0 \\ q_i(x,0) &= 0 \\ uC_i - D\frac{\partial C_i}{\partial x}\bigg|_{x=0} &= uC_i^0 \\ \frac{\partial C_i}{\partial x}\bigg|_{(L,t)} &= 0 \end{aligned}$$

After setting parameter values to those listed in Table 2, choosing the Langmuir isotherm from table 5 and isotherm parameters from figure 13, a solution to the system was computed (Figures 5 and 6).

3 Tools used for resolution

Membranes have their own specific characteristics, and experimental results from the adsorbate-adsorbent interaction have lead to different adsorption models (e.g., those included in Tables 4 and 5). These models describe the equilibrium kinetics in adsorbate-adsorbent interactions. Nonetheless, many adsoprtion isotherm models may similarly fit a set of q_i^{eq} data as a function of C_i^{eq}. For example, Langmuir's isotherm is similar to Hill's when $n = 1$ in the latter, and similar to Redlich-Peterson when $g = 1$.

Table 4: One-component-one-site isotherms.

Isotherm	Equation	Unknown Parameters*
Langmuir	$q^{\mathrm{eq}} = \frac{Q_{\mathrm{o}} b C_{\mathrm{e}}}{1+bC_{\mathrm{e}}}$	Q_{o} (mg/mL), b (mL/mg)
Freundlich	$q^{\mathrm{eq}} = K_{\mathrm{F}} C_{\mathrm{e}}^{1/n}$	K_{F} (mg/mL)$^{1-\frac{1}{n}}$, n (unitless)
Dubinin-Radushkevich	$q^{\mathrm{eq}} = Q_{\mathrm{o}} \mathrm{e}^{-K_{\mathrm{ad}}\epsilon^2}$, where $\epsilon = RT \ln\left(1 + \frac{1}{C_{\mathrm{e}} \times 10^3}\right)$	Q_{o} (mg/mL), K_{ad} (mol/J)^2, ϵ (J/mol)^2
Temkin	$q^{\mathrm{eq}} = \frac{RT}{b_{\mathrm{T}}} \ln\left(A_{\mathrm{T}} C_{\mathrm{e}}\right)$	b_{T} (mol-mL/J-mg), A_{T} (mL/mg)
Flory-Huggins	$\frac{\theta}{C_{\mathrm{o}}} = K_{\mathrm{FH}} (1-\theta)^{n_{\mathrm{FH}}}$, where $\theta = \frac{q^{\mathrm{eq}}}{Q_{\mathrm{o}}}$ and $C_{\mathrm{o}} = \frac{C_{\mathrm{e}}}{1-\theta}$	Q_{o} (mg/mL), θ (unitless), K_{FH} (mL/mg), n_{FH} (unitless)
Hill	$q^{\mathrm{eq}} = \frac{Q_{\mathrm{o}} C_{\mathrm{e}}^{n_{\mathrm{H}}}}{K_{\mathrm{D}} + C_{\mathrm{e}}^{n_{\mathrm{H}}}}$	Q_{o} (mg/mL), K_{D} (mg/mL)$^{n_{\mathrm{H}}}$, n_{H} (unitless)
Redlich-Peterson	$q^{\mathrm{eq}} = \frac{K_{\mathrm{R}} C_{\mathrm{e}}}{1+a_{\mathrm{R}} C_{\mathrm{e}}^{g}}$	a_{R} (mg/mL)$^{-g}$, K_{R} (unitless), g (unitless)
Sips	$q^{\mathrm{eq}} = \frac{K_{\mathrm{s}} C_{\mathrm{e}}^{\beta_{\mathrm{S}}}}{1+a_{\mathrm{S}} C_{\mathrm{e}}^{\beta_{\mathrm{S}}}}$	a_{S} (mg/mL)$^{-\beta_{\mathrm{S}}}$, K_{S} (mg/mL)$^{1-\beta_{\mathrm{S}}}$, β_{S} (unitless)
Toth	$q^{\mathrm{eq}} = \frac{K_{\mathrm{T}} C_{\mathrm{e}}}{(a_{\mathrm{T}} + C_{\mathrm{e}})^{1/t}}$	a_{T} (mg/mL), t (unitless), K_{T} (mg/mL)$^{1/t}$
Koble-Corrigan	$q^{\mathrm{eq}} = \frac{A C_{\mathrm{e}}^{n}}{1+B C_{\mathrm{e}}^{n}}$	A (mg/mL)$^{1-n}$, B (mg/mL)$^{-n}$ n (unitless)
Khan	$q^{\mathrm{eq}} = \frac{q_{\mathrm{s}} b_{\mathrm{K}} C_{\mathrm{e}}}{1+b_{\mathrm{K}} C_{\mathrm{e}}^{a_{\mathrm{K}}}}$	q_{s} (mg/mL)$^{a_{\mathrm{K}}}$, b_{K} (mg/mL)$^{-a_{\mathrm{K}}}$ a_{K} (unitless)
Radke-Prausnitz	$q^{\mathrm{eq}} = \frac{a_{\mathrm{RP}} r_{\mathrm{R}} C_{\mathrm{e}}^{\beta_{\mathrm{R}}}}{a_{\mathrm{RP}} + r_{\mathrm{R}} C_{\mathrm{e}}^{\beta_{\mathrm{R}}-1}}$	a_{RP} (unitless), r_{R} (mg/mL)$^{1-\beta_{\mathrm{R}}}$, β_{R} (unitless)

*Parameter units are consistent with our simulations, where C_{e} and q^{eq} are in mg/mL. If present, R is the gas constant (8.3145 J/mol-K) and T is temperature (298.15 K). Refer to [5] for additional isotherm and parameter description.

Table 5: Multicomponent-one-site isotherms.

Isotherm	Equation	Unknown Parameters*
Langmuir	$q_i^{\mathrm{eq}} = \frac{Q_{\mathrm{o},i} b_i C_{\mathrm{e},i}}{1+\sum_{j=1}^{n} b_j C_{\mathrm{e},j}}$	$Q_{\mathrm{o},i}$ (mg/mg), b_i and b_j (mL/mg),
Freundlich	$q_i^{\mathrm{eq}} = K_{\mathrm{F},i} C_{i,\mathrm{e}} \left(\sum_{j=1}^{n} a_{ij} C_{\mathrm{e},j}\right)^{n_i - 1}$	$K_{\mathrm{F},i}$ (mg/mL)$^{1-n_i}$, a_{ij} (unitless), n_i (unitless)
Khan	$q_i^{\mathrm{eq}} = q_{\mathrm{s},i} b_{\mathrm{K},i} C_{i,\mathrm{e}} \left(b_{\mathrm{o}} + \sum_{j=1}^{n} b_{\mathrm{K},j} C_{\mathrm{e},j}\right)^{a_{\mathrm{K},i}}$	$q_{\mathrm{s},i}$ (mg/mg), $b_{\mathrm{K},i}$ (mL/mg) b_{o} (unitless), $a_{\mathrm{K},i}$ (unitless)
Fritz-Schlunder	$q_i^{\mathrm{eq}} = \frac{a_{i,\mathrm{o}} C_{i,\mathrm{e}}^{B_{i,\mathrm{o}}}}{b_i + \sum_{j=1}^{n} a_{ij} C_{\mathrm{e},j}^{B_{ij}}}$	$a_{i,\mathrm{o}}$ (mg/mg)(mg/mL)$^{-B_{i,\mathrm{o}}}$, a_{ij} (mg/mL)$^{-B_{ij}}$, $B_{i,\mathrm{o}}$ (unitless), B_{ij} (unitless), b_i (unitless)

*Parameter units are consistent with our simulations, where C_{e} and q^{eq} are in mg/mL and mg/mg, respectively. Refer to [6] for isotherm and parameter description. Subscripts i and j represent solute.

3.1 Isotherm evaluation (q_{e}, the direct problem)

Given the equilibrium concentration of the adsorbate (C_{e}) and the isotherm model parameters, the direct problem computes the amount of adsorbate in the absorbent at equilibrium (q_{e}) by using the different single-

component and multi-component isotherm models in Tables 4 and 5, respectively.

3.2 Isotherm optimization problem (the inverse problem)

The optimization problem is based on nonlinear least-squares approaches to obtain model parameters that best fit the experimental data. It is expressed as follows:

$$\min_{\mathbf{p}\in\Omega\subseteq\mathbb{R}^m} \|\mathbf{q}_{\text{exp}} - \mathbf{q}_{\text{model}}(\mathbf{p})\|_2^2 , \tag{14}$$

where $||.||_2^2$ is the squared Euclidean norm; $\mathbf{q}_{\text{exp}}$ is the vector that contains the experimental (or simulated) amount of adsorbate in the adsorbent at equilibrium; $\mathbf{q}_{\text{model}}$ is the isotherm model version of $\mathbf{q}_{\text{exp}}$; $\mathbf{p}$ is the vector with m real elements (isotherm model parameters) to be obtained from the optimization routine; and Ω is the subset of the m-dimensional space that encloses $\mathbf{p}$. The set Ω is restricted by the bounds of each element of vector $\mathbf{p}$ which, from now on, will be referred to as the lower and upper bound vectors, denoted as $\mathbf{p}_{\text{l}}$ and $\mathbf{p}_{\text{u}}$, respectively.

3.3 Numerical Methods

To optimize Equation (14), we developed a global optimization algorithm that combines two methods: simulated annealing, which is a stochastic algorithm that finds satisfactory solutions in problems analogously to a physical annealing process [8], and a descent method, based on Newton's method, that considers bounds on the variables [3].

3.4 Simulated annealing method

Simulated Annealing (SA) is a technique that mixes combinatorial optimization with statistical mechanics [7]. It is a mechanism that tries to find satisfactory solutions rather than optimal solutions, although in some cases they may coincide. Contrary to the descent method, the SA is a stochastic method that does not require information about the gradient and Hessian matrix of the objective function.

The relation between the SA method with statistical mechanics derives from its analogy of finding a set of parameters that best minimizes a function with the arrangement of particles in a physical system with the lowest internal energy. From this analogy, the arrangement of particles or physical state is the parameter vector $\mathbf{p}$ that is to be obtained from the optimization routine, and the internal energy is the optimization function in Equation (14) evaluated at $\mathbf{p}$.

As with physical systems, SA consists of melting the system (at sufficiently high temperatures to explore the parameter space) and cooling it (decrease the temperature) slowly, to avoid a bad annealing process, but not too slowly to avoid a large run time. In addition to this melting-cooling process, the SA explores the vicinity of a current state to see if a lower energy state is found.

Briefly, the main components of the SA algorithm are:

1. *The Cauchy annealing schedule.* It assures a fast annealing at higher temperatures and a slow one at lower temperatures. It is defined as

$$T_{i+1} = \frac{T_i}{i+1}, \tag{15}$$

   where T_i and T_{i+1} are the current and new temperatures, respectively, and i is the "external" iteration, or the iteration where temperature is varied.

   The initial and final temperatures depend on the range scale where the optimized function is. They also depend on the desired run time of the optimization routine. High initial or low final temperature values can lead to longer run times. On the other hand, low initial temperature values can lead to a vague exploration of the parameter space and high final temperatures can lead to an incomplete annealing process, i.e., the "best" minimum is not found.

2. *The Cauchy distribution as the new candidate generator when the temperature is constant.* It permits occasional long jumps in the current neighborhood. These new candidates are obtained as

$$\mathbf{p}_{r+1} = \mathbf{p}_r + \gamma \tan\left[(\theta - \frac{1}{2})\pi\right], \tag{16}$$

where $\mathbf{p}_{r+1}$ and $\mathbf{p}_r$ are the new and current parameter vectors (equivalent to the new and current physical states); γ is the location parameter vector whose entries are defined as

$$\gamma_k = 10^{\log p_{r,k}},\ k = 1, 2, ..., m, \tag{17}$$

where m is the size of vector $\mathbf{p}_r$; and θ is a vector whose entries are between zero and one and are selected randomly with uniform distribution. This candidate generator permits occasional long jumps in the current neighborhood as $\theta_k \to 0$ or $\theta_k \to 1$ for any k.

3. *The uniform distribution as the new candidate generator when the temperature changes.* This is

$$\mathbf{p}_{r+1} = \mathbf{p}_r + \alpha^{\mathrm{T}} \left(\mathbf{p}_{\mathrm{u}} - \mathbf{p}_{\mathrm{l}} \right), \tag{18}$$

where $\alpha \in \mathbb{R}^m$ is uniform and each element is in the range $[0, 1]$.

4. *The acceptance and transition probabilities.* These are: the probability of moving from one state to another in the same neighborhood at constant temperature ($\mathrm{P_a}$) and the probability of moving to higher energy states when the temperature changes ($\mathrm{P_t}$). In SA, the new candidate is accepted or rejected depending on the Boltzmann probability distribution, defined as

$$P = \mathrm{e}^{\frac{\Delta f_{r+1}}{T}} > P_x,\ r = 1, 2, ..., R-1, \tag{19}$$

where $\Delta F_{r+1} = F_r - F_{r+1}$; R is the global number of iterations; T is temperature; P_x is the minimum probability at which a move at constant temperature is accepted ($x = \mathrm{a}$) or the minimum probability at which a transition to another neighborhood is made ($x = \mathrm{t}$). If $\Delta f > 0$, then $P_x = 1$.

Note from Equation (19) that as $T \to \infty$, all the states occur with probability equal to one ($\mathrm{e}^{\frac{\Delta f}{T}} \to 1$) making it possible to move to different configurations where the internal energy is higher. On the contrary, as $T \to 0$, only the states of lower energy will have nonzero probability ($\mathrm{e}^{\frac{\Delta f}{T}} \to 0$ if $\Delta f < 0$).

The selection of lower energy states is favored as the acceptance probability P_{a} tends to one. In addition, a small value of P_{t} permits the exploration of the surface, particularly at high temperatures, preventing the algorithm to get stuck at a local minima. Moreover, as the system "cools", it is less possible to move to another neighborhood, unless the new state has lower energy than the current state.

Therefore, the acceptance and transition probabilities were selected to assure that the best solution inside a neighborhood is found and to permit the exploration of areas that are not necessarily inside a neighborhood of the current parameter vector, regardless that the new state is not one of lower energy.

Fixed parameters for the SA routine are included in Table 6.

Table 6: Simulated annealing algorithm conditions that are kept fixed during the optimization routine.

Condition	**Description**	**Value**
$\mathrm{T_{initial}}$	Initial temperature	0.1
$\mathrm{T_{final}}$	Final temperature	0.001
$\mathrm{j_{max}}$	Maximum number of iterations	100
$\mathrm{P_a}$	Acceptance probability	0.9
$\mathrm{P_t}$	Transition probability	0.2

Global optimization solutions for problems similar to that in Equation (14) can sometimes be difficult to achieve, since usually the shape of the parameter surface where the function is optimized is unknown. Therefore, SA increases the probability of exploring a large part of, if not all, the studied surface. However, since SA finds satisfactory solutions rather than optimal solutions, we then use a descent method to optimize the selected SA parameter.

Table 7: Algorithm conditions for the descent method.

Condition	Value
Termination tolerance of $\mathbf{p}$	10^{-8}
Termination tolerance of $F(\mathbf{p})$	10^{-8}
Maximum number of iterations	10^{6}
Maximum number of function evaluations	10^{6}

3.5 Descent method

The selected descent method (DM) is a technique based on Newton's method that is used for the optimization of large-scale nonlinear problems subject to bounds in some of their variables [3]. It keeps the solution within the bounded region.

Let f be equal to the optimization function defined in (14), then the optimization problem can be written as

$$\min_{\mathbf{p}\in\Omega\subseteq\mathbb{R}^m} f(\mathbf{p}), \tag{20}$$

where m is the size of $\mathbf{p}$.

Coleman and Li in [3] used a technique that replaces f by its quadratic approximation using a Taylor expansion. In their technique, the variables (the elements of vector $\mathbf{p}$) are constrained to the region by varying the radius of a ball that encloses the variables. This ball is scaled with respect to the region bounds.

At the n^{th} iteration the resultant sub-problem can be written as follows:

$$\min_{\mathbf{s}_n\in\mathbb{R}^m} \left\{ \mathbf{g}_n^{\mathrm{T}}\mathbf{s}_n + \frac{1}{2}\mathbf{s}_n^{\mathrm{T}} H_n \mathbf{s}_n, \quad ||D_n^{-1}\mathbf{s}_n|| \leq \Delta_n \right\}, \tag{21}$$

where $\mathbf{s}_n = \mathbf{p}_{n+1} - \mathbf{p}_n$; $\mathbf{g}_n$ is the gradient of $f(\mathbf{p}_n)$, and H_n is the Hessian matrix of $f(\mathbf{p}_n)$; $\Delta_n > 0$ changes at each iteration n; and D_n is the n^{th} iteration of the diagonal scaling matrix D defined as

$$D_n(\mathbf{p}) = \mathrm{diag}(|\mathbf{v}(\mathbf{p})|^{1/2}), \tag{22}$$

where

$$v_i = \begin{cases} p_i - u_i, & \text{if } g_i < 0 \\ p_i - l_i, & \text{if } g_i \geq 0 \end{cases}$$

for $i = 1, 2, ..., m$; p_i is the i^{th} entry of $\mathbf{p}$; and l_i and u_i are elements of the region bounds, which are the entries of vectors $\mathbf{p}_{\text{l}}$ and $\mathbf{p}_{\text{u}}$, respectively.

The scaling matrix D prevents a step directly toward a boundary point and it assures that the solution of the problem will be found inside the bounds. To choose D in (22) Coleman and Li in [3] considered the optimality conditions of the constrained optimization problem in (20).

To implement the DM, we used the Matlab® function for large-scale optimization of nonlinear least-squares subject to bounds, `lsqnonlin`. Given the bounded region and an initial vector $\mathbf{p}$ within the bounded region, which was initially obtained with the SA algorithm (Figure 7), the `lsqnonlin` function finds an optimum by iteratively solving the objective function.

The algorithm conditions for the DM are included in Table 7. From this table, the parameter termination tolerance allowed is the maximum parameter vector difference in two consecutive iterations, $||\mathbf{p}_{n+1} - \mathbf{p}_n||$, and the function termination tolerance allowed is the maximum function difference between two consecutive iterations, $|f(\mathbf{p}_{n+1}) - f(\mathbf{p}_n)|$.

3.6 Optimization Process

Separation processes are complicated and feature several mathematical challenges. Designing columns to meet our goals is nontrivial, as each need is different and requires a different set of design factors. We can, through optimization, recommend a membrane-based separation setup which is most beneficial to its user. To perform this optimization, we used a Genetic Algorithm (GA). A GA is an evolutionary algorithm which

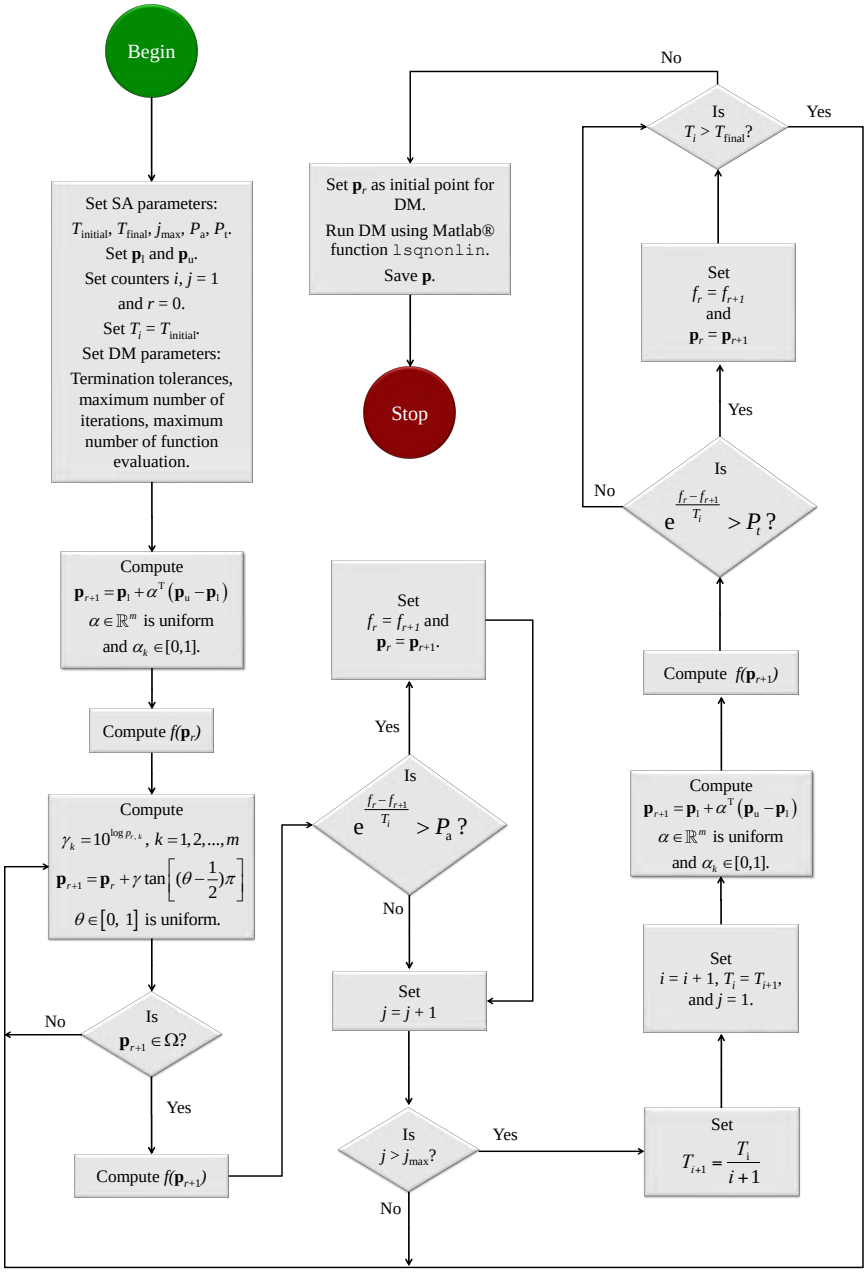


Figure 7: Simulated annealing-descent method (SA-DM) algorithm.

is popular in engineering optimization. It mimics the principles of genetics and the Darwinian principle of natural selection. With a GA, the design points represent individuals in a population of solutions. Objective functions provide fitness ratings for each individual, and we sort the population by fitness. At each step, the GA selects at random from the highest-ranked (most fit) individuals in the current population to be parents and uses them to produce the children for the next generation. Thus, we see the parallel between a GA and "survival of the fittest." Over successive generations, the population "evolves" toward a solution [13][14]. While there is no convergence theory to prove that a GA converges to a global optimum, it is still used in many fields and is respected for its versatility. The GA is sometimes criticized for requiring a large function budget, but for our application this is not an issue, as all computations are performed quickly.

We use the NSGA-II (Non-dominating Sorting Genetic Algorithm) developed by the Kanpur Genetic Algorithms Laboratory, and available in MATLAB from Mathworks' File Exchange. The NSGA-II operates on real-valued variables and has built-in bound constraint handling. Inequality or equality constraints must be handled using outside methods, such as applying a penalty-barrier method. The NSGA-II features the simulated binary crossover (SBX) and polynomial mutation as its genetic operators.

There are many decision variables which need to be considered when performing optimization, such as injection velocity, injection concentration, filter length, porosity, and time. Our objective functions also differ according to different applications, such as minimizing cost or maximizing yield. Different constraints also may need to be considered according to different applications.

4 Objective Functions

We examine two distinct applications of membrane-based filtration methods. Each application has different goals, so we separate them and design objective functions for them based on individual needs. First, we examine a water purification application, where we design a column for removing one or more contaminants from water with the goal of creating potable water. In this problem, our goals are to increase the lifetime of the filter while ensuring we produce as much potable water as we can. In the second scenario, we examine the removal of a valued protein from a solution by membrane-based filtration. In this case, the primary goal is to recover as much protein as possible.

4.1 Application 1: Purification

We seek to maximize the volume of purified liquid. This is achieved by maximizing the time until breakthrough, which we write as

$$\max f(u_{inj}, C_{in}, L, \theta_0) = t_b \tag{23}$$

where u_{inj} is the fluid injection rate and thus the flow rate of the fluid through the column, C_{in} is the concentration of the contaminant at the inflow, L is the length of the column, and θ_0 describes the initial porosity of the filter.

For this application, we define the breakthrough time t_b by

$$t_b = \min\left\{t : \frac{C(L,t)}{C_{in}} > 0.95\right\} \tag{24}$$

where $C(L,t)$ is the concentration of the contaminant at the right boundary at time t. Thus breakthrough time is, effectively, the time after which the filter stops being acceptably effective at removing contaminants. For this application, we want the filter to last as long as it can, to minimize replacement or cleaning requirements.

We also seek to maximize the volume of purified water that we produce. We approximate the volume of purified water produced by considering the injection flow rate (here, assumed constant) and the time until breakthrough. As a result, our objective function is

$$\max g(u_{inj}, C_{in}, L, \theta_0) = ut_b \tag{25}$$

Reducing power consumption would reduce overall costs. As a result, we seek to minimize the power required to filter water using this filter column. A large power sink in this process is the pump which pushes

water through the filter. This pump must apply pressure to the water, pushing it through the filter and, ultimately, out of the column. As the filter becomes clogged by debris, our assumption of constant flow rate results in increased pressure required to push water through the column.

To model this effect, we use the equation for relating porosity as a function of adsorbed ions given in [9].

$$\ln\left[\frac{1-\theta(t_i)}{1-\theta(t_0)}\right] = \eta(Q)Q(t_i) \tag{26}$$

where $\theta(t)$ is the porosity at time t (note that, here, $\theta(t_0) = \theta_0$ is a decision variable), and $\eta(Q)$ is the filter's adsorption coefficient with total adsorption $Q(t)$ at time t. We compute $\eta(t)$ by creating a cubic spline using existing data points from (Madaeni, Salehi, '09). We evaluate the resulting polynomial to approximate $\eta(Q(t_b))$.

We rearrange eq. (26) to solve for $\theta(t_i)$, giving us

$$\theta(t_i) = 1 - e^{\eta Q(t_i)}(1-\theta(t_0)) \tag{27}$$

Once we have solved for $\theta(t_i)$, we solve the Kozeny-Carman equation

$$k(\theta) = \frac{\theta^3}{5S_0^2(1-\theta)^2}I \tag{28}$$

where S_0 is the sphericity of particles (assumed to be 1), and I is defined by

$$I = \frac{D_p^2}{36L} \tag{29}$$

where D_p is the diameter of the spherical particle, and L is the length of the column. We can then write our objective function for minimizing power consumption as

$$\min h(u_{inj}, C_{in}, L, \theta_0) = \Delta p = \frac{\mu u_{inj}}{k(\theta)} \tag{30}$$

where μ is the viscosity of the fluid. As we are simulating water filtration at or near standard temperature and pressure, we know that μ is equal to 8.9×10^{-4} Pa·s.

The NSGAII is designed to minimize objective functions. Thus, to maximize objectives f and g, we must reformulate them as minimization objectives. This is simply done by forming the objectives

$$\min f_2 = \frac{1}{1+f} \tag{31}$$

and

$$\min g_2 = \frac{1}{1+g} \tag{32}$$

This results in objectives which, when minimized, will provide solutions which maximize our original objective functions f and g.

4.2 Application 2: Pharmacology

In this part, we'll consider a biological application, which requires the extraction of protein from a certain solute. First, we'll consider the one-component model. With this model, we considered two general objective functions. The first, which seeks to maximize the volume of recovered biological product, is given by

$$\max J(u_{inj}, C_{in}, L, \theta_0) = (1-\theta)Q \tag{33}$$

where

$$Q = \int_0^{t_b}\int_{col} q\,dx\,dt \tag{34}$$

for a single-component system. For an n-component system, we have

$$Q = \int_0^{t_b} \int_{col} \sum_{i=1}^{n} q_i \, dx \, dt \tag{35}$$

Here, t_b is given by

$$t_b = \min \left\{ t : \frac{C(L,t)}{C_{in}} > 0.10 \right\} \tag{36}$$

As in Application 1, we must reformulate our objective J to be a minimization objective. The objective becomes

$$\min J_1 = \frac{1}{1+J} \tag{37}$$

Minimizing this objective will give solutions that maximize the original objective, as desired.

Our second objective seeks to minimize the pressure drop required to remove the desired protein and push the solution through the column.

$$\min J_2 = \Delta P = g(u_{inj}, C_{in}, L, \theta_0, t_{final})$$

where u_{inj} is injection velocity, C_{in} is injection concentration, L is column length,θ_0 is the initial porosity of the filter,and t_{final} is the maximum filtering time. ΔP is the increased pressure required to push through the column as previously described. Here, we want to maximize the yield while minimizing the cost, which is represented by ΔP. We use this to represent cost because increasing the pressure drop increases the mechanical power required to maintain flow and thus the energy consumption, which is costly.
For the two-component model, different scenarios are considered. Suppose we have two components C_1 and C_2 in the solute. According to different applications, we have

- Case 1: for a certain solute, in which the ratio of C_1 and C_2 concentration is constant, we want to maximize the product of S1 while minimizing C_2.
- Case 2: the concentration ratio of C_1 and C_2 can be adjusted, and we want to maximize the production of C_1 while minimizing the cost.

For Case 1 of the two-component model, the objective functions are

$$\min J_1 = 1/(1+Y_1) = f(u_{inj}, L, \theta_0, t_{final})$$
$$\min J_2 = Y_2 = f(u_{inj}, L, \theta_0, t_{final})$$

where Y_1 and Y_2 are the yield of C_1 and C_2. Other variables are the same as before.

5 Constraints

5.1 Bound Constraints

An important feature of the NSGAII is that is allows the enforcement of bound constraints on all decision variables. This allows us to ensure that nowhere in the optimization does any decision variable leave a reasonable range of values. The lower and upper bounds presented in Table 8 represent the bounds used for our optimization runs. Note, θ_0 is intentionally left unitless.

We also considered applying a bound constraint on the amount of purified water produced in Application 1. This would be designed around a minimal water output to sustain those drinking the water. If too little water is produced, they will suffer, so maintaining at least a lower bound of water production is important. This is accomplished by applying an additive penalty scaled by constraint violation. That is, we construct

$$\mathcal{P}(g_2, K) = g_2 + K(u, t_b) \tag{38}$$

Variable	Lower Bound	Upper Bound
u_{inj}	30 cm/min	60 cm/min
C_{in}	0.1 g/mL	10 g/mL
L	5 cm	100 cm
θ_0	0.5	1

Table 8: This table contains upper and Lower bounds for decision variables, used by the NSGAII for optimization bounds.

where

$$K(u, t_b) = \begin{cases} \rho(V_{min} - ut_b) & \text{if } ut_b < V_{min} \\ 0 & \text{otherwise} \end{cases} \tag{39}$$

Here V_{min} denotes the minimum desired volume of purified water, and ρ is a penalty weight parameter. When the constraint is violated, the additive penalty results in a function value that indicates a nonideal solution. However, we must be aware that the parameter ρ affects the effectiveness of this penalty. If ρ is too low, then a constraint violation may not be apparent. If it is too high, it will dominate otherwise appealing solutions which may be salvageable through the genetic operations of the NSGAII. A generally-accepted estimate for ρ is a value on the same order of magnitude as expected function values. In this example, we used a value of 10 for ρ, as no objective exceeded this order of magnitude.

Then, we seek to minimize $P(g_2, K)$. If no constrains are violated, this is identical to minimizing g_2. This sort of additive penalty typically helps the optimizer to move design points out of the infeasible space and can prevent new design points from entering infeasible space.

6 Numerical results

Here, we show verification for the transport model we constructed by comparing to an analytical solution. We also examine the efficiency of the stiff and non-stiff nonlinear solvers available and select from them the best for our application. Finally, we perform optimization and analysis using modeled isotherms within our transport equations. This is done because different systems are governed by different isotherms. As such, we need a transport model capable of handling them.

6.1 Error analysis for verification

In this part we try to confirm that the code solves the equations with the expected convergence behavior for the methods we are using. To accomplish this, an analytical solution was found for the initial and boundary conditions as $C(x, 0) = 0, \frac{\partial C}{\partial x}(L, t) = 0, C(0, t) = c_0(t), q(x, 0) = 0$. Parameters were also set to be: $\theta = 0.5, u = 1, D = 1, k_m = 2, L = 1, c_0(t) = t$, with

$$q^{eq} = -0.5\cos(\pi x) + t\pi\sin(\pi x) - t\pi^2\cos(\pi x) + q,$$

and the exact solutions as the following:

$$q = -t\cos(\pi x) + t^2\pi\sin(\pi x) - t^2\pi^2\cos(\pi x),$$

$$c = t\cos(\pi x)$$

. The default tolerance settings for the "ode15s" integrator in Matlab–relative tolerance as 1e-3 and absolute tolerance as 1e-6, and a half step size–were used to obtained the following result (Table 9). The final concentration value was also used to compute the convergence errors. It is shown that the convergence order with respect to Δx, the spatial step size, is first order.
As for the error for q, it stays below the error tolerance after the first time step, so we do not show it here.

Table 9: Error and Convergence order of the concentration C

$\Delta x = 0.1$	**Error**	**Convergence Order**
Δx	4.16e-02	1.00e+00
$\Delta x/2$	2.43e-02	7.76e-01
$\Delta x/4$	1.30e-02	9.00e-01
$\Delta x/8$	6.73e-03	9.53e-01
$\Delta x/16$	3.42e-03	9.77e-01
$\Delta x/32$	1.72e-03	9.89e-01
$\Delta x/64$	8.65e-04	9.94e-01

6.2 Efficiency Comparison

In this section we compare the computational efficiency of the following three matlab time integrators– ode15s, ode23s and ode45–for our system. We do this with the aim of determining the most efficient solver for our system, meaning that with the same accuracy demand it will need the least amount of time. Though our problem has just one spatial dimension, we will need to solve it for every function call in the optimizer, which implies that it may well be the bottleneck for the performance of the whole system. Thus it is not only necessary but also crucial to guarantee that we select best solver for this system. Generally we do this by fixing the spatial step size as a relatively small size first. Then we reduce the error tolerance for the time quadrature and check how much time the respective solves require.

Tables 10–12 compare the accuracy for the three integrators as the time integration tolerances were refined from 10^{-2} to 10^{-6} Each table shows the computational time required, the infinity norm at the final time step, and the number of time steps required. In addition to the total computational time, the number of time steps is a good parameter to illustrate the cost for reaching a particular accuracy. For convenience of the comparison, we also provide a graph with total computational time as a horizontal coordinate and the log of tolerance as a vertical coordinate. In the graph, we only show ode15s and ode23s since the table clearly shows that ode45 does not perform as well since it is an explicit formula. From the graph we can see that ode15s shows a higher efficiency compared with ode23s.

Table 10: ODE15s Temporal Error

CPU Time	L_∞ Error	Time Steps Required
6.00e-02	6.86e-03	12
8.00e-02	8.32e-04	14
8.00e-02	7.62e-05	17
9.00e-02	6.99e-06	21
8.00e-02	8.15e-07	27

Table 11: ODE23s Temporal Error

CPU Time	L_∞ Error	Time Steps Required
3.40e-01	5.57e-04	11
4.10e-01	2.76e-04	12
6.30e-01	9.15e-05	20
1.43e+00	5.01e-05	41
3.16e+00	6.86e-06	88

Table 12: ODE45 Temporal Error

CPU Time	L_∞ Error	Time Steps Required
3.59e+00	3.80e-02	8757
3.97e+00	1.93e-03	8781
3.71e+00	1.06e-04	8781
3.77e+00	1.90e-05	8789
3.65e+00	8.47e-06	8793

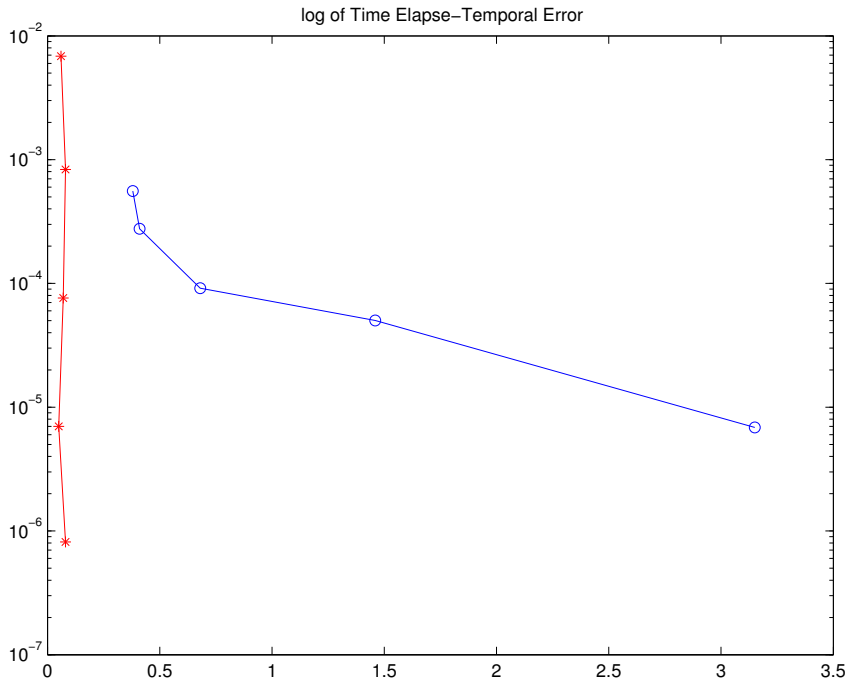


Figure 8: Comparison of ode15s (red) and ode23s (blue)

6.3 Isotherm Fits

Datasets from [10] and [1] were used as for validation of the isotherm models. [10] contained equilibrium data for the two adsorbate concentrations of protein and sodium chloride (NaCl), but q_e data are only available for the adsorption of protein, as these were the data of interest. Thus, only parameters relevant to calculating q_e for the protein were regressed in this case. [1] studied the adsorption of chromium(VI) and nickel(II) at pH values of 1.0 and 4.5, and provided equilibrium data for the adsorption of both compounds. A table of the raw experimental data was not available for [1], so figure analysis software (Datathief) was used to extract the data from the figures; this process introduced slight additional uncertainty in the experimental data.

Based on the residual sum of squares (RSS), the Bayesian Information Criterion (BIC) was calculated [11] to add a penalty factor for the number of parameters fit to the data. A higher BIC is evidence against the optimality of a model.

$$\mathrm{BIC} = n \ln(\mathrm{RSS}/n) + k \ln(n) \tag{40}$$

where RSS is the residual sum of squares, n is the number of data points, and k is the number of parameters estimated to fit the data.

The isotherm parameters and fitting statistics for the one-site isotherm models are presented in Tables 13 and 14, respectively. The results for the two-component model suggest that the Khan model with five fitting parameters provides the best fit for the experimental data, although the Langmuir model provides a reasonable fit with fewer parameters.

Fit isotherm parameters and statistics for the [1] data are presented in Tables 15 and 16, respectively. The first three models give acceptable fits, but the empirical model proved to provide the best fit of the data. The predictions are plotted against the experimental data in Figure 10. The calculated BIC parameters provide evidence that this model is superior to the others, even after adding a penalty for the additional fitting parameters available.

Table 13: One-component- and two-component-one-site models

Model Type	Isotherm	Parameter	Value
One-component, one-site	Langmuir	Q_{o}	1
		b_1	13.52
	Hill	Q_{o}	1
		n_{H}	2
		K_{D}	0.00547
Two-component, one-site	Langmuir	$Q_{\mathrm{o},1}$	0.204
		b_1	2.05
		b_2	9.09
	Freundlich	$K_{\mathrm{F},1}$	4.98
		a_{11}	175
		a_{12}	805
		n_1	0.287
	Khan	$b_{\mathrm{K,o}}$	43.7
		$b_{\mathrm{K},1}$	30.6
		$b_{\mathrm{K},2}$	127
		$q_{\mathrm{s},1}$	2.37
		$a_{\mathrm{K},1}$	-1.45
	Fritz & Schlunder	a1,o	3.09
		a_{11}	1.11
		a_{12}	53.9
		b_1	23.6
		$B_{1,\mathrm{o}}$	0.360
		B_{11}	-1.55
		B_{12}	1.09

Units are consistent with those presented in Tables 4 and 5.

Table 14: Fitting results for 2-component-2-site isotherm models

Model	# parameters	SSR	R^2	BIC
Langmuir	3	1.23E-03	0.975	-228
Freundlich	4	4.32E-03	0.913	-194
Khan	5	6.85E-04	0.986	-232
Fritz & Schlunder	7	1.99E-03	0.960	-203

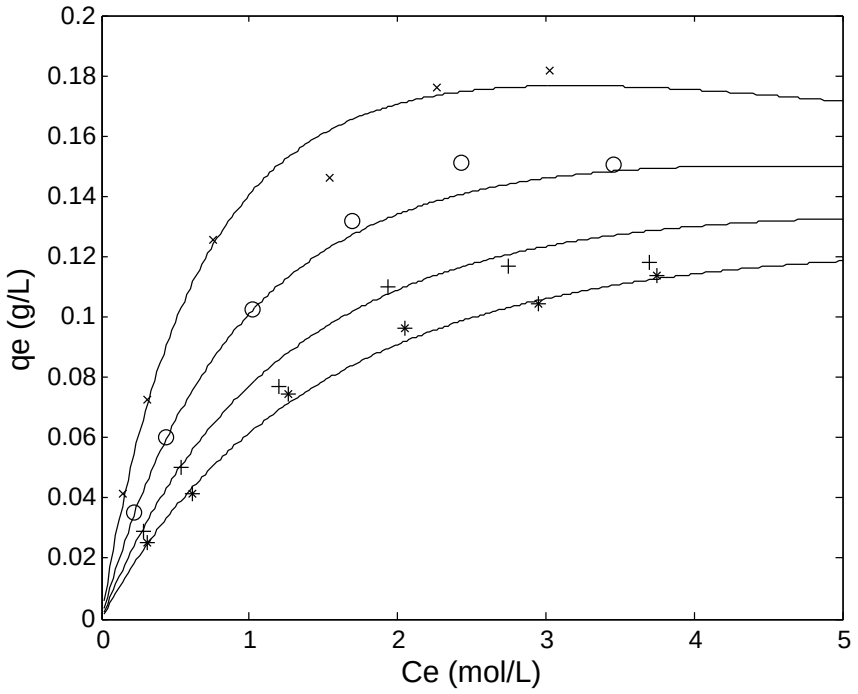


Figure 9: Khan isotherm model vs. data for protein adsorption. Concentrations of NaCl: x 0 mol/L, o 0.15 mol/L, + 0.30 mol/L, * 0.45 mol/L.

Table 15: Isotherm fit parameters for chromium and nickel in activated sludge at pH 4.5. Fit parameters were obtained using adsorbate concentrations of mg/L and adsorbent concentrations of mg/g.

Langmuir			
$Q_{o,1}$	$Q_{o,2}$	b_1	b_2
175	287	1.61E-03	3.63E-03

Fruendlich							
$K_{F,1}$	$K_{F,2}$	a_{11}	a_{12}	a_{21}	a_{22}	n_1	n_2
45.6	66.3	332	166	103	293	0.514	0.565

Khan						
$b_{K,0}$	$b_{K,1}$	$b_{K,2}$	$q_{s,1}$	$q_{s,2}$	$\alpha_{K,1}$	$\alpha_{K,2}$
50	1.46	3.08	2.38	4.33	-0.471	-0.483

Fritz & Schlunder						
$a_{1,o}$	$a_{2,o}$	a_{11}	a_{12}	a_{21}	a_{22}	b_1
24.3	95.8	100	0.0693	2.93	7.30	4.78
b_2	$B_{1,o}$	$B_{2,o}$	B_{11}	B_{12}	B_{21}	B_{22}
76.6	0.437	1.26	-0.898	0.565	0.688	0.820

6.4 Application 1: Purification

6.4.1 One-Site Constant Porosity Model

This problem involves the minimization of three objective functions, with only bound constraints enforced natively by the NSGAII. The optimal design point found for this model was $x^* = (35.2775,\ 49.9999,\ 31.3455,0.5003)$. This solution gives function values $(f_2, g_2, h) = (0.2185,.0076,18.6056)$. This equates to approximately a 4.6 minute breakthrough time, 130.5 mL of purified water produced, and a pressure drop of 18.6 kPa required. Pareto fronts for each pair of objective functions are given below, in Figures 6.4.1, 6.4.1, and 6.4.1.

Table 16: Isotherm fit statistics for chromium and nickel adsorption suggest that the Fritz and Schlunder model provides the best fit for these data.

Model	p	SSR	$(R^2)_1$	$(R^2)_2$	BIC
Langmuir	4	1663	0.931	0.995	243
Freundlich	8	537	0.988	0.997	179
Khan	7	1018	0.956	0.997	221
Fritz & Schlunder	14	290	0.993	0.998	160

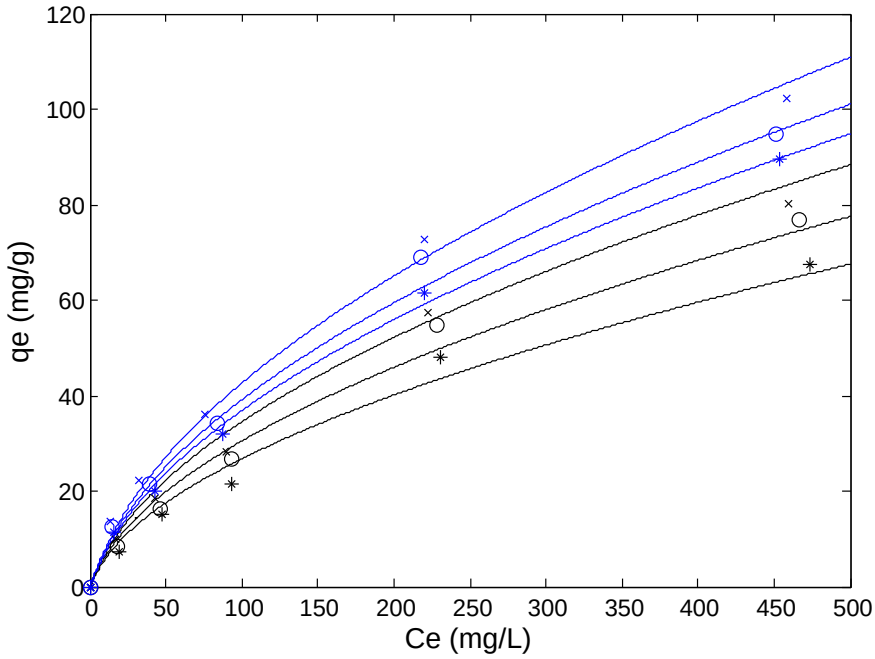


Figure 10: Fits for chromium adsorption at pH 4.5. Initial concentration nickel (mg/L): x 0, o 25, * 50, x 100, o 250, * 500.

We see that there is an inverse relationship between ΔP and t_b. This is expected, as an increased pressure drop will cause the filter to become less effective more rapidly, decreasing breakthrough time. Note the gaps in each Pareto front; there are areas where the GA did not select an points. This suggests that design points which generate these solutions are not within our bound sonstraints. It is also possible that such solutions cause poorly-conditioned matrices within the model, or result in high error in the use of ode15s. This is a difficulty that can arise when we evaluate design points - some design points have high error when ode15s attempts to solve the nonlinear system of equations. However, such design points are weeded out by the GA as they produce solutions with significantly higher function values than those selected the genetic operations.

To better understand our model, we performed a sensitivity analysis on each of the objectives with respect to each of the decision variables. We generated 4000 points using the built-in MATLAB function `lhsdesign` to create a latin hypercube sampling of our design space. This is, essentially, a random set of design points which are approximately uniformly distributed through the design space. We then checked each objective for significant design variables and for interactions between them. The results can be seen in Table 6.4.1.

These analyses tell us that u_{inj} has a significant effect on every objective, which is reasonable. In fact, it is the only significant decision variable for minimizing ΔP. This is particularly useful if design specifications are known, so L can be fixed, or if C_{in} is controllable. Reducing the number of decision variables allows us to make better optimization decisions with less effort. It is also interesting to see that L has no significant effect on volume produced or power consumption. This means that we can fix L and get optimal solutions for the

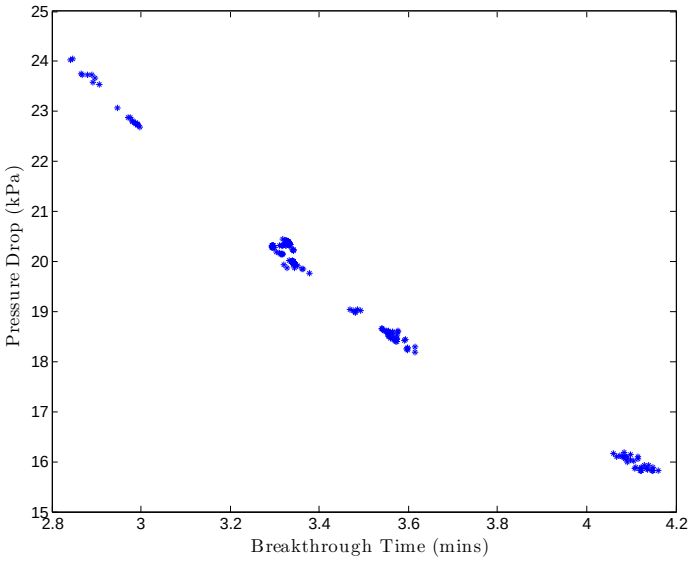


Figure 11: This Pareto front shows the inverse relation between Δ P and t_B. We also notice the gaps along the front. These suggest that the design points which result in solutions in those areas are not within the bounds we used.

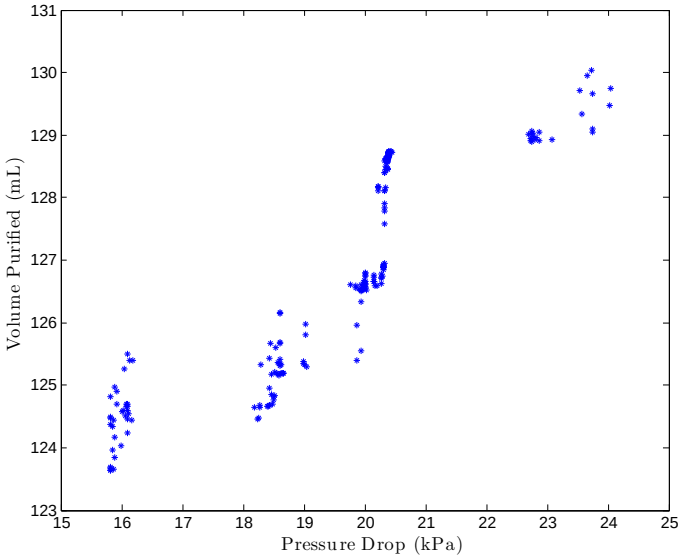


Figure 12: This Pareto front shows that as we increase the pressure drop (Δ P) we are able to purify a greater volume before breakthrough occurs. This is because a greater applied pressure will force water through the column more rapidly.

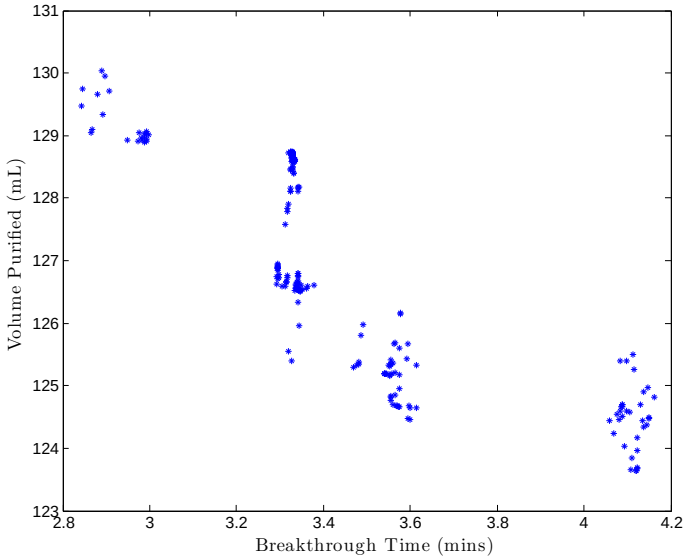


Figure 13: This Pareto front shows that as breakhtrough time t_b is increased, we obtain a lower volume of purified water. This occurs because as we process more water, we cause the filter to be less effective, until eventual breakthrough.

Breakthrough Time		
Variable	**P-value**	**Significant?**
u_{inj}	1.96×10^{-92}	Yes
C_{in}	0	Yes
L	0.0445	Yes
θ	9.45×10^{-15}	Yes
Volume		
Variable	**P-value**	**Significant?**
u_{inj}	0.008	Yes
C_{in}	0	Yes
L	0.118	No
θ	2.30×10^{-21}	Yes
Δ P		
Variable	**P-value**	**Significant?**
u_{inj}	0	Yes
C_{in}	0.179	No
L	0.359	No
θ	0.096	No

Table 17: ANOVA table for one-site adsorption model for water purification. Significance is at level $\alpha = 0.05$. P-values of zero are due to MATLAB's limitation in representing values smaller than 10^{-323}.

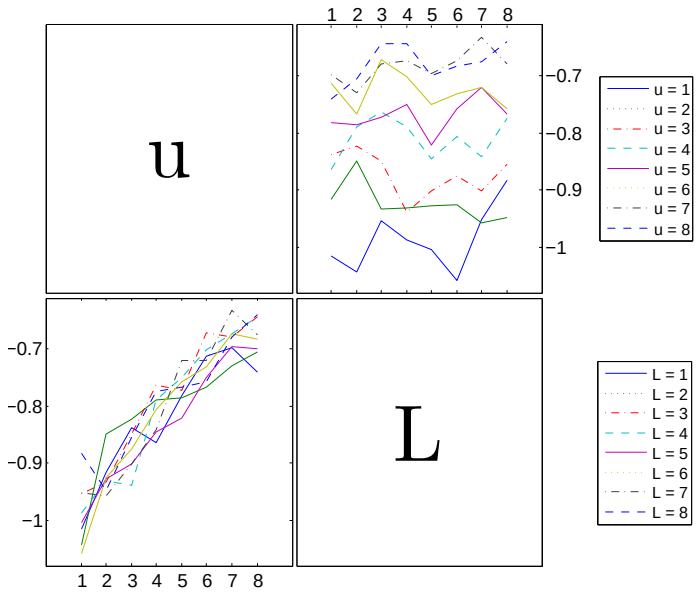


Figure 14: Interaction plot for u and L with respect to breakthrough time objective. The upper-right plot shows main effects (vertical axis) at different levels of u as L changes (horizontal axis). The lower-left plot shows main effects (vertical axis) at different levels of L as u changes (horizontal axis).

remaining factors if we desire. This is good because column size is an easily-controllable factor which plays into the cost and ease-of-use of the column.

We also examine interactions between the significant main effects. The most notable interactions we see are under the Breakthrough objective, between u and L. This interaction is given by Figure 6.4.1. We see that as we change L for a given level of u, the objective function value changes significantly. The converse is also true. Similarly, we observe an interaction between u and C_{in}. For minimizing power, we see several interactions. C_{in} interacts with L and θ. L interacts with θ and u, and u interacts with L. We did not investigate higher-order interactions. However, despite seeing many interactions in these objectives, we observe no interaction effects with respect to volume purified.

6.4.2 Two-Site Variable Porosity (Clogging) Model

The previous results apply only when the porosity of the filter does not change as contaminants are captured by the filter. This is unrealistic, as in our filtration system the pores will become clogged with bound contaminants as the filtration progresses. As such, we need to include in the model the changes in porosity that occur during filtration, as represented by eq. (27). We still treat θ_0 as a decision variable, but within the model itself, porosity can vary over time. We also consider two separate binding sites within the filter. These sites may be designed to capture different elements of the fluid, or may simply trap using different mechanisms, or some combination. We examined the same objective functions, this time featuring two binding sites. This leads to a more realistic model, and should prompt more useful solutions. The primary modeling difference between this case and the one-site case is that this case calculates Q using eq (35) whereas the single-site case calculates Q using eq (34). Additionally, our decision variables must be altered to differentiate between the injection concentrations of our two subspecies. As a result, C_{in} becomes $C_{in,1}$ and $C_{in,2}$.

Table 18: Design Variables Setting

variable	min	max
u_{inj}	0.01	0.1
C_{in}	0	1
L	1	50
θ_0	0.1	0.9
t_{final}	1	3600

Optimization of this case resulted in the solution

$$(u_{inj}, C_{in,1}, C_{in,2}, L, \theta_0) = (60, 2.3205, 9.9468, 88.585, 0.14286)$$

This solution results in function values $(f_2, g_2, h) = (0.0174, 0.0277, 8.54)$. This suggests a breakthrough time of 56.48 minutes, a total of 3037 mL of purified water, and a pressure drop of 8.54 kPa. This is approximately 53.75 mL of purified water per minute, which is sensible given our injection rate of 60 mL per minute. Interestingly, this solution lies on the boundary condition for u_{inj}, implying that the solver wanted to pump water through faster, which would likely result in increased volume purified. However, such a change would also likely increase the pressure drop required to pump and decrease breakthrough time.

6.5 Application 2: Pharmacology

In this part, we'll consider a biological application, which needs to extract protein from certain solute. First, we'll consider the 1-component model. With the 1-component model, we considered two general objective functions.

$$\min J_1 = 1/(1+Y) = f(u_{inj}, C_{in}, L, \theta_0, t_{final})$$
$$\min J_2 = |\Delta p| = f(u_{inj}, C_{in}, L, \theta_0, t_{final})$$

where Y is the yield,u_{inj} is injection velocity, C_{in} is injection concentration, L is column length,θ_0 is the initial porosity of the filter,and t_{final} is the filtering time. p is the increased pressure requried to push through the column as described before. Here, we want to maximize the yield while minimize the cost which is represented by $|\Delta p|$.

For the one-component model, we chose the injection velocity, concentration, filter length, initial porosity of the filer, and the filtering time according to previous research and possible choices [14]. The range of these variables are listed in Table 18. As the Pareto front shows in Figure 15, the yield increases as the pressure change increases. When the pressure change reaches a certain level, the yield doesn't change significantly beyond that.

Because the property of isotherms is very important to the filtering process as we discussed before, we also add the coefficients of isotherms to our design variables based on above optimization process. For the linear isotherm case, we take the upper and lower bounds of the coefficients to be -1 and 1 in order to increase computing efficiency, and the Pareto front figure is shown in Figure 16. Different from the above results, the pressure change is much smaller, which indicates the importance of isotherm property for the separation process. Similarly as previous results, the yield changes more at the beginning and less as the increase of pressure change. For Case 1 of the two-component model, we set the concentration of S1 and S2 both to be 0.5 for the experiment. And we choose our design variables similar as before (Table 19). Similarly, the linear isotherm is added to the design variables, and we take the upper and lower bounds of the coefficients both to be -1 and 1 in order to increase computing efficiency. The result is shown in Figure 18.

For Case 2 of the two-component model, we also choose our design variables similar as before(Table 20). The yield of Y1 increases more at the beginning of pressure change, while it does not vary much with the increase of pressure change after a certain value. Similarly, we inlcude the coefficients of linear isotherms in optimization process, the result is shown as Figure 20. The yield of Y1 increases dramatically at the beginning of pressure change while doesn't change much as the increase of pressure change after a certain value.

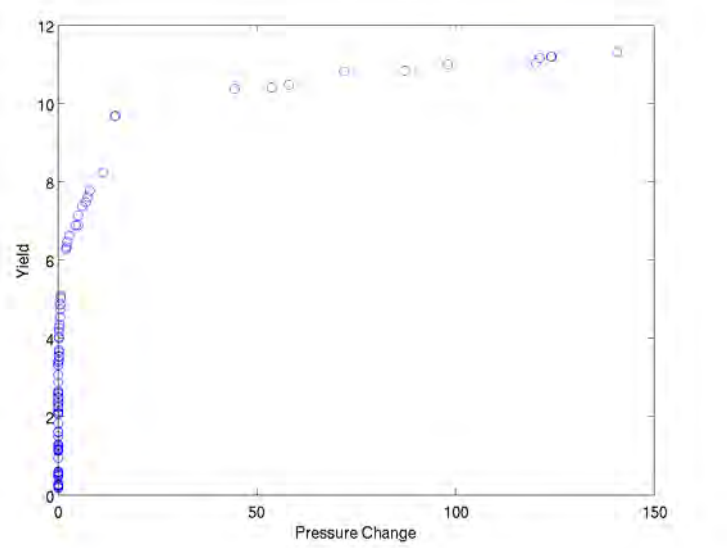


Figure 15: Pareto front for Pressure Change and Yield of 1-component model with fixed isotherm coefficients

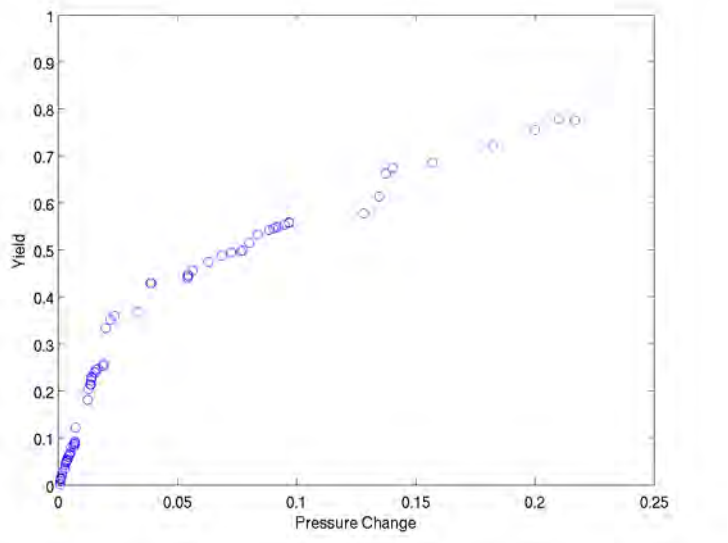


Figure 16: Pareto front for Pressure Change and Yield of 1-component model with linear isotherm coefficients in design variables

Table 19: Design Variables Setting for Case 1 of 2-component model

variable	min	max
u_{inj}	0.01	0.1
L	1	10
θ_0	0.1	0.9
t_{final}	1	3600

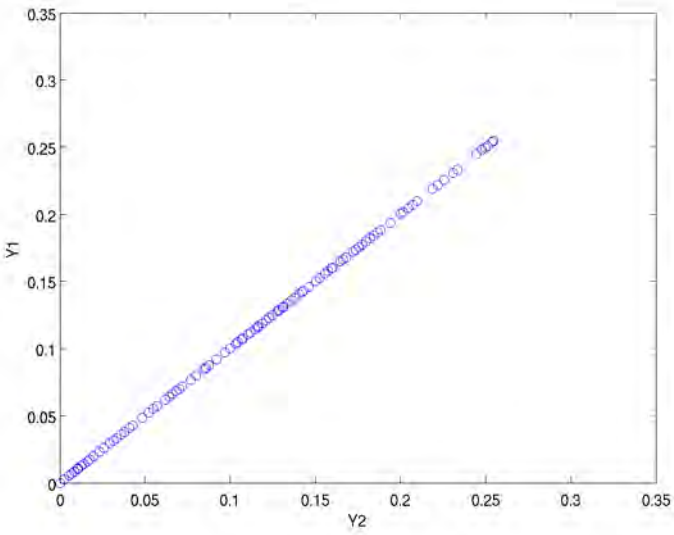


Figure 17: Pareto front for S1 and S2 Yield for Case 1 of 2-component model

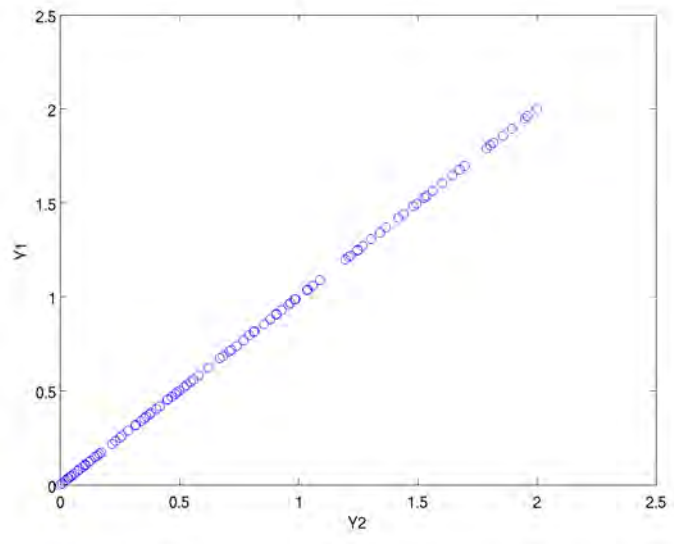


Figure 18: Pareto front for S1 and S2 Yield for Case 1 of 2-component model with isotherm coefficients in design variables

Table 20: Design Variables Setting for Case 2 of 2-component model

variable	min	max
u_{inj}	0.01	0.1
L	1	50
C_{in1}	0	1
C_{in2}	0	1
θ_0	0.1	0.9
t_{final}	1	60

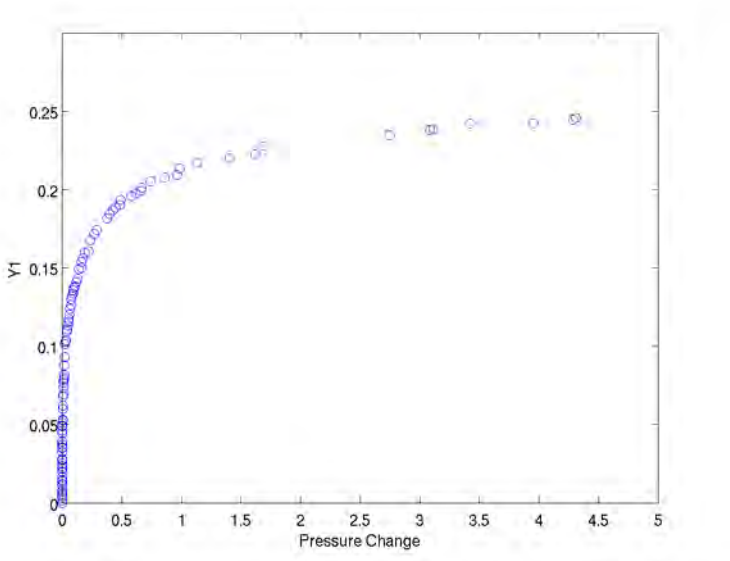


Figure 19: Pareto front for S1 and S2 Yield for Case 2 of 2-component model

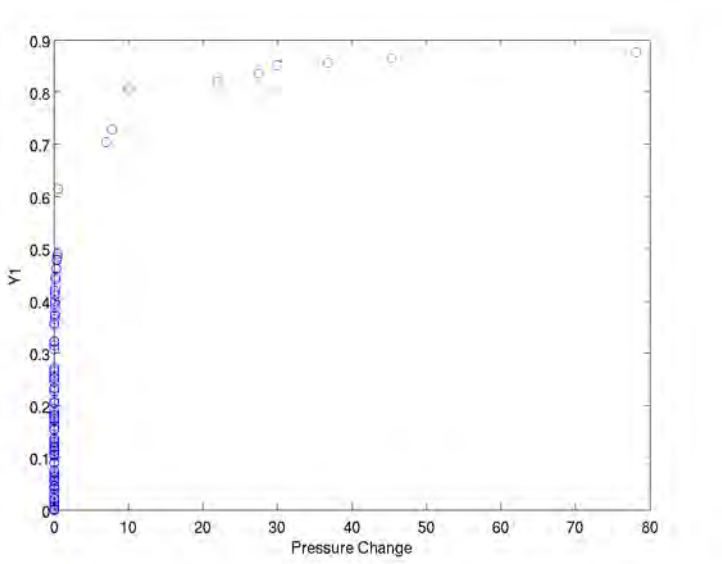


Figure 20: Pareto front for S1 and S2 Yield for Case 2 of 2-component model with isotherms coefficients in design variables

7 Conclusions

Mathematical modeling is a very powerful technique, and it can be very useful to formulate hypotheses of system responses under different perturbations. Furthermore, it can help us in the development of new techniques and/or equipment prior testing them *in bench*, minimizing research efforts, costs, and time. Constant mathematical modeling leads to improvement of algorithms to obtain more accurate results. Moreover, since mathematics is the universal language of Science, the theoretical concepts used to build these mathematical models for water purification via membrane separation, e.g., fluid dynamics, chemical kinetics, and transport phenomena, can also be applied to other, e.g., oceanography and physiological systems.

Further research is needed due to the complexity of separation processes. In this project, only the 1-D model is considered. In reality, 2-D and 3-D models are needed to better represent the dynamics of separation processes. Additionally, multi-component and multi-site models can be touched on in the future, which can provide more information and guidance for industry applications. Accordingly, different isotherms and optimization methods can be explored to further improve the representation of separation processes and computing efficiency.

The current setup for performing optimization could be modified. We are currently using the NSGA-II, but we could alter our optimization process to solve the more complex cases of the problem (multiple components, multiple dimensional flow) more rapidly and with greater precision. We could consider other algorithms; a local search algorithm may be an effective choice when we have a better grasp of initial iterates based on experimental data. We could consider using derivative-based methods featuring a numerical Hessian or approximated gradient. These can, if properly constructed, save considerable computation time. Rather than requiring thousands of function evaluations, some derivative-based algorithms can find a solution within n^2 evaluations, where n is the number of decision variables. Finally, we could consider hybridization. Hybridization involves combining a global search with a local search to form one algorithm. This allows the algorithms to build on each other, playing to the strengths of both and avoiding their weaknesses.

We could also perform more statistical analysis. We only analyzed the one-component constant porosity model. In an experimental setting, we are more likely to confront multiple-component systems with clogging. If we want to understand these problems more deeply, we have to perform a statistical analysis of them as well. We could potentially gain a lot of understanding of the evolution of our system if we were to perform sensitivity analysis. This is especially useful for reducing the number of decision variables to a more manageable number.

7.1 Recommendations for industry

Optimal water purification processes require complex equipments. Nonetheless, mathematical models could be key in determining what should be the characteristics of these equipments, e.g., their dimensions and geometry, to maximize efficiency in the processes and minimize costs and time. For example, let us suppose that a certain industry wants to improve its water purification and product recovery processes. However, they do not want to change the adsorption membrane they are using for two reasons: first, producing it is highly expensive and, second, small-scale testing with this membrane has shown very appropriate kinetics for the product(s) they want to recover. In addition, they mentioned that the actual equipment was constructed from a "large-scale" adaptation of the one they used to obtain membrane-desired product kinetics. During this small-scale-to-large-scale adaptation, they ended up with low product recovery and high costs due to high pressure drops. In addition, they ended up with several by-products that resulted from large retaining times and the membrane breakthrough time was very high.

Based on the results presented in this report, where we observed an inverse relation between pressure drop and breakthrough time, we suggest that there may be additional chemical reactions going on in the separation system as result of the large column retaining times. This appears to be having an effect on the amount of recovered product. By adding mathematical models that describe the chemical kinetics of this by-product production in the overall water purification/product recovery model, the equipment may be optimized so that product recovery increases, breakthrough time decreases, pressure drop decreases, leading to cost minimization and profit maximization.

This is just a simple example of how mathematical models, like the one we propose, could lead to possible solutions on how to optimize equipment for water purification/product recovery. Moreover, this analysis can be applied if the industry has the equipment, but needs to design an optimal membrane for water purification/product recovery.

References

[1] Z. Aksu, U. Acikel, E. Kabasakal, and S. Tezer. Equilibrium modelling of individual and simultaneous biosorption of chromium(vi) and nickel(ii) onto dried activated sludge. *Water Research*, 36:3063–3073, 2002.

[2] MAJ William J. Bettin. Water purifiers for the warfighter. *Army Medical Department Journal*, 2007.

[3] T.F. Coleman and Y. Li. An Interior Trust Region Approach for Nonlinear Minimization Subject to Bounds. *SIAM J Optimiz*, 6(2):418–445, 1996.

[4] M.W. Farthing, C.E. Kees, T.F. Russell, and C.T. Miller. An ELLAM approximation for advective-dispersive transport with nonlinear sorption. *Advances in Water Resources*, 29:657–675, 2006.

[5] K.Y. Foo and B.H. Hameed. Insights into the modeling of adsorption isotherm systems. *Chem Eng J*, 156(4598):210, 2010.

[6] A.R. Khan, T.A. Al-Bahri, and A. Al-Haddad. Adsorption of phenol based organic pollutants on activated carbon from multi-component dilute aqueous solutions. *Wat Res*, 31(8):2102–2112, 1997.

[7] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[8] M. Locatelli. Simulated Annealing Algorithms for Continuous Global Optimization. In *Handbook of Global Optimization II*. Kluwer Academic Publishers, 2002.

[9] S.S. Madaeni and E. Salehi. A new adsorption and porosity combined model for passage of cations through nanofiltration membrane. *Journal of Membrane Science*, 303:100–109, 2009.

[10] B.K. Nfor, M. Noverraz, S. Chilamkurthi, P.D.E.M. Verhaert, L.A.M. van der Wielen, and M. Ottens. High-throughput isotherm determination and thermodynamic modeling of protein adsorption on mixed mode adsorbents. *Journal of Chromotography A*, 1217:6829–6850, 2010.

[11] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[12] S.-Y. Suen and M.R. Etzel. A mathematical analysis of affinity membranes bioseparations. *Chemical Engineering Science*, 47(6):1355–1364, 1992.

[13] A. Tarafder. Modeling and multi-objective optimization of a chromatographic system. In Gade Pandu Rangaiah and Adrin Bonilla-Petriciolet, editors, *Multi-Objective Optimization in Chemical Engineering*, pages 369–398. John Wiley & Sons Ltd, 2013.

[14] C. C. Yuen, Aatmeeyata, S. K. Gupta, and A. K. Ray. Multi-objective optimization of membrane separation modules using genetic algorithm. *Journal of Membrane Science*, 176(2):177–196, August 2000.

Geographic and Racial Differences of Persons Living with HIV in the Southern United States

Tyler Massaro[1], Christina Edholm[2], Rachel Grotheer[3], Isabel Chen[4], Yiqiang Zheng[5]

Faculty Mentors: Simone Gray[6], Howard Chang[7]

Abstract

African Americans are almost three times as likely as whites to be living with HIV in the southern United States. In addition, socioeconomic indicators such as income, education attained, and employment status, have been linked to HIV prevalence. This study sought to quantify the contribution of race and socioeconomic determinants to the overall presence of HIV in the southern United States. We collected 2010 U.S. Census data including non-hispanic white proportion (NHW), non-hispanic black proportion (NHB), Hispanic proportion, education attainment level, poverty level, urban status, income, and unemployment rate for 1,422 counties in the 16 states represented in the U.S. South. We then performed three types of regression modeling (multiple linear, conditional autoregressive, Bayesian Poisson hierarchical model), non-metric multidimensional scaling, and two types of cluster analysis (K-Means, Besag-Newell). The K-Means cluster analysis was used to cluster counties without using any geographic information, while the Besag-Newell considers the geographic layout of the counties. The regression analyses showed robust association between HIV prevalence and several demographic variables. Specifically, a county-level 1% increase in NHB percentage was associated with a 2.5% (95% C.I. 2.1% – 2.9%) increase in HIV prevalence accounting for socioeconomic variables and unmeasured spatial confounders. The non-metric multidimensional scaling, comparing social determinants of health across the 16 states, showed proximity between Maryland and Delaware, Tennessee and Georgia, while Texas was an outlying observation. The K-Means analysis failed to identify a cluster of counties whose HIV prevalence rate was significantly different from the regional HIV rate. By comparison, the Besag-Newell clustering solution contained counties whose HIV rate was significantly different (p-value <0.001) from the regional rate. When visualized on a map, we saw that these clusters corresponded to large metropolitan areas, including Memphis, Miami, Atlanta, D.C., and Baltimore. Using these counties in the regression model, we determined that NHB remained the most important indicator for HIV prevalence rate in the southern United States.

[1]Mathematics, The University of Tennessee, Knoxville
[2]Mathematics, University of Nebraska-Lincoln
[3]Mathematical Sciences, Clemson University
[4]Mathematics, Emory University
[5]Mathematics and Statistics, Purdue University
[6]Centers for Disease Control & Prevention
[7]Emory University

1 Introduction

Since the HIV outbreak began in the United States in the 1980s, the disease has spread to all corners of the country, and results in an estimated 50,000 new cases each year [16]. The epidemic has demonstrated a higher prevalence in African Americans [3, 2, 4, 14, 26, 32], and persons with low income [13]. In comparison to other regions in the United States, both of these subpopulations are over-represented in the South [6, 13, 14, 32, 37]. Not coincidentally, the disease affects residents in the southern states at a higher rate than in other parts of the country [5, 6, 14, 26, 30, 32, 34, 37].

Evidence shows that individual, social and structural factors have been linked to the increased mortality and morbidity of diseases [2, 4, 6, 9, 13, 14, 26]. These determinants of health include educational attainment, unemployment, access to healthcare and poverty rates. African Americans are disproportionately affected by these social and economic barriers making them more susceptible to most diseases [1, 4, 5, 12, 21, 26, 29, 32, 34, 37, 46]. Of the non-hispanic Black population (NHB), 26.49% of reside in areas below the poverty line, compared to 12.11% of non-hispanic whites [39]. Similarly, 17.93% of NHB in the United States have less than a high school level education, compared to just 12.15% of NHW [39].

African Americans are almost three times as likely as whites to have HIV in the southern United States [39]. HIV incidence is also higher in regions of low income and lower education [13, 32, 33, 35], as well as in areas of higher urbanicity and higher unemployment [13, 14]. Unemployment rates in Florida are more than three times as high as those in North Dakota; the proportion of people living with HIV in Florida is more than 18 times that of North Dakota. Similarly, in Texas the percent of the population without high school education is almost 4 times as high as North Dakota [8, 39]; the proportion of people living with HIV in Texas is almost 10 times that of North Dakota [8, 39].

To date, most national and state-wide studies of the relationship between HIV and at-risk populations have been conducted at the state level. As more data has been collected and made accessible, efforts have been made to perform healthcare research on a finer scale, specifically at the county level [12, 17, 20, 29, 31, 35, 38]. Studies show that results at this level have the potential to be more useful to local healthcare professionals[35].

This study focuses on understanding the relationship between living with HIV in the southern United States and social determinants of health (SDH). We use county-level surveillance and census data to evaluate this relationship. We aim to enhance HIV literature by providing a novel evaluation of the various social determinants and their contribution to HIV prevalence at the county level. This information will help influence policy decisions at the county level, and provide opportunities for improving community health.

2 Methods

2.1 Data

2.1.1 HIV Surveillance Data

HIV surveillance data were obtained from the National Center for HIV/AIDS, Viral Hepatitis, STD and TB Prevention (NCHHSTP) Atlas at the U.S. Centers for Disease Control and Prevention (CDC). The NCHHSTP Atlas, provided county level estimates of annual prevalence rates (per 100,000) and people living with diagnosed HIV in 2010[8]. The Atlas data also included information on race/ethnicity. The data were restricted to the southern United States as defined by the U.S. Census bureau [41] as the following 16 states: Alabama, Arkansas, Delaware, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia and West Virginia.

In order to ensure privacy, data were suppressed for counties with population less than 100or fewer than 5 cases of people living with HIV [8]. This suppression resulted in missing data for approximately 12% of the counties in the U.S. South. In order to gain an overall picture of people living with HIV in the U.S. South and to facilitate spatial analysis, we imputed the number of cases of people living with diagnosed HIV for each data suppressed county. We assumed that the counties with data suppressed had only 0-4 cases of people living with HIV. Since there was only one county (Loving County, TX) with population less than 100, this assumption was reasonable for all imputation criteria.

To impute missing prevalence rates, first we calculated state-specific average prevalence rates using the total number of reported cases and the total population size among reporting counties. Assuming the missing counties shared the same state-wide prevalence, we then calculated the expected number of cases in the missing counties based on population estimates from the 5-year American Community Survey (ACS) for the period 2008-2012 (<http://factfinder2.census.gov> accessed July 2014). If the imputed number of cases was greater than 4, we replaced the value with 2, since we assumed all the data suppressed counties had fewer than 5 cases. We chose 2 since it is the median value between 0 and 4. If the imputed number of cases was less than 4 we kept the imputed value for the number of cases. From the imputed number of cases we calculated the imputed rate using the population of the county.

2.1.2 Demographic Data

We obtained data for several demographic variables from the 2008-2012 ACS 5-Year Estimates using 2010 census data available at (<http://factfinder2.census.gov> accessed July 2014). We used the 5-year estimates since it is the most reliable source of information at the county level and encompasses the study period [42].

Each variable was chosen based on the possibility of a cause-effect relationship between the parameter and the number of people living with HIV in the U.S. South. These include unemployment rate, percent of the population with less than a high school education, percent of the population below the poverty line, median household income, urbanicity, percent of the population identified as non-Hispanic Black/African American (NHB), percent identified as Hispanic/Latino, and percent identified as non-Hispanic White (NHW). The urbanicity indicator assigned counties a value of 1 if they had at least 50,000 inhabitants. This classification is based on the US census definition of an urbanized area [40].

2.2 Statistical analyses

The statistical analysis consists of two major parts. The first part focused on the disease clustering to spot out the high HIV prevalence areas and associated regression analysis within clusters. The later part was devoted to investigate the possible associations between HIV prevalence and racial and socioeconomic variables using regression methods.

2.2.1 Besag-Newell clustering analysis

We begin by detecting clusters of high-risk areas using only regional HIV prevalence data. We define a cluster to be a sub-region that has a higher relative risk of disease incidence than its neighbors. There are two questions we aim to address: are there clusters in the region of interest, and if so, where are they located? Once clusters are identified, we perform regression analysis within the clusters to determine the effect of social, economic and racial composition on disease incidence in these high risk areas.

It is important to bear in mind that in this context, only geographical location and prevalence counts are taken into account. None of the SDH indicators are used to determine whether a specific area is considered a cluster. To make this distinction clear, we will use the term *spatial clusters*.

Spatial relationships between the regions are described by a matrix of weights, denoted W. Often, W is a measure of adjacency, and takes the form

$$w_{ij} = \begin{cases} 1, & \text{if region } i \text{ and region } j \text{ share a boundary, } i \neq j \\ 0, & \text{else.} \end{cases}$$

Another approach is to use the distances between regions as a weight, and so W may take the form

$$w_{ij} = d_{ij},$$

or some other function of the distances d_{ij} between the regions (see, for example, Tango's EET method [36]).

In this paper we look at two global measures of clustering, both of which are adjacency-based, and will therefore have the first form of W appearing in the chosen test statistic. We compute a statistic developed by P.A.P Moran (Moran's I index) in 1950 [28], and another by R.C. Geary (Geary's contiguity ratio c) in 1954 [15]. The only other data required to compute both test statistics are the population size, the prevalence counts, and the expected incidence counts for each region. The test statistics computed will give an indication of the presence, or lack, of spatial clustering. That is, they assess whether rates for neighboring areas are more similar than would be expected if they were randomly distributed among the geographic areas. These computations, however, do not indicate where the clusters are located.

To identify the locations and extent of clusters, we implement a cluster detection method developed by Besag and Newell [7]. The goal is to identify regions of higher relative risk, based solely on the underlying geographical structure of the region of interest. The null hypothesis under consideration is that of a spatially homogeneous relative risk. Suppose that the study region is divided into n small areas indexed by $i = 1, \ldots, n$ with observed and expected (under the null) number of cases denoted by O_i and E_i. Let P_i be the total number of persons at risk in region i and λ the annual rate of cases for the region of study. Under H_0, we assume that the incidence rate for all individuals is constant in all regions, therefore the expected number of cases is proportional to the size of the population in a given region. Explicitly,

$$E_i = \frac{O_+}{P_+} P_i$$

where O_+ is the total incidence count over the entire region of study, and P_+ is the total at risk population. The null model H_0 assumes that the observed incidence counts O_i are independently distributed as Poisson random variables with mean $\lambda = E_i$.

In this paper, we do not adjust for confounding factors such as sex and age, because the HIV data were not be stratified at these levels. We used the implementations provided in the R package 'DCluster'.

2.2.2 Global Measures of Clustering

We considered 2 measures of global spatial dependence. Moran's I index is a measure of spatial autocorrelation in which the similarity between regions is defined as the product of the respective deviations:

$$I = \frac{\sum_i \sum_j w_{ij}(r_i - \bar{r})(r_j - \bar{r})}{w_{..} \sum_i (r_i - \bar{r})^2 / n}$$

Here r_i is the incidence rate of region i. When neighboring regions tend to have similar rates, I will be positive. When neighboring regions tend to have different rates, I will be negative. Values range from -1 (indicating perfect dispersion) to $+1$ (perfect correlation). A zero value indicates a random spatial pattern.

The second approach is uses Geary's contiguity ratio c, which is the ratio of the sum of mean squared differences between rates for all pairs of areas:

$$c = \frac{n-1}{n} \frac{\sum_i \sum_j w_{ij}(r_i - r_j)^2}{w_{..} \sum_i (r_i - \bar{r})^2 / n}$$

If rates are geographically distributed at random, the contiguity ratio is close to one. Low values less than one indicate positive autocorrelation.

2.2.3 Besag & Newell, 1991

We then applied Besag and Newell's [7] method to obtain estimates of the likely clusters' geographical location and extent based solely on the incidence of disease at each region. This method considers windows with a pre-fixed number of cases k, the size of the cluster. Each region with nonzero cases is considered in turn as the center of a possible cluster. For each such region, the smallest number of neighbouring regions S needed to reach k cases is computed. A small observed value of S indicates that there is a cluster centered at the starting region. The p-value of each potential cluster is given by

$$\mathbb{P}(S \leq s) = 1 - \sum_{t=0}^{k-1} \exp\left(-u_s Q)(u_s Q)^t\right)/t!$$

where $Q = O_+/P_+$ and u_s is the total population within the cumulated region S. If the observed p-value is small enough, the region and its surrounding S neighbours will be considered a significant cluster.

The implementation we used was from the R package SpatialEpi. Possible clusters are centered at the centroids of each county. Distances between counties are also computed using the respective centroids.

The clustering approach contains 2 parameters that need to be pre-defined: (1) the size k of the cluster, and (2) the significance level α. Typically α is chosen to be much less than the usual 0.05 or 0.01 levels. We chose $\alpha = 0.05/C$, where C is the total number of counties in the U.S. South. This resulted in a very small significance level, which unfortunately results in a high Type II probability error, but this was necessary to avoid multiple testing.

A priori determination of the cluster size k is a major drawback of this method. To deal with the arbitrariness of the choice of parameter k, we ran the test using 8 different values of k. There was quite some variation in the locations and sizes of clusters obtained for different values of k. To deal with this variability, we defined

the clusters to be *the counties that were detected by the Besag-Newell method for multiple values of k*. In this way, we incorporate the results across a range of k values and thereby reduce the influence of a single arbitrary choice of k on the results. The choices for k were determined by the distribution of incidence counts over the region. We chose 8 k values that spanned the median (29) and the maximum (25, 564) incidence counts across the U.S. South. Specifically, we worked with $k = 50, 100, 300, 500, 1000, 15000, 20000, 25000$.

At the region-wide level, preliminary analysis showed a positive relationship between k and the number of clusters detected[8]. For some values of k, we obtained close to 500 clusters, many of which overlapped to some degree. To reduce computational intensity, for each k value, we selected only the most significant 9 clusters for further analysis. However, because the observed p-values were so small (in the order of 10^{-19}), one can no longer ensure the correct ranking of clusters based on p-values. We therefore also considered the standardized mortality ratio ($SMR_i = O_i/E_i$) as another indicator of significance. Out of the identified clusters, we selected only those that have SMR exceeding an a priori threshold of 2. From these, we selected 9 clusters with the smallest p-value.

2.2.4 Regression Analysis on Spatial Clusters

Simple linear regression analysis was performed on the counties belonging to clusters identified by the Besag-Newell method. We evaluated the effect of each predictor variable on the log-transformed HIV prevalence rate, and compared this to the regional results. For each of these counties, we also fitted the regression model

$$\log(\text{HIVRate}) \sim \%\text{Less than HS} + \text{Med income} + \text{Urb. Ind.} + \text{NHB} + \text{NHW}$$

2.2.5 Multidimensional scaling and K-means cluster analysis

Multidimensional scaling (MDS) provides a method for understanding patterns in observations as opposed to variables [23]. In particular, it is useful for visualizing similarities between observations arising in a high-dimensional setting. We used non-metric MDS in two ways while conducting this research.

First, we performed state-level MDS in the southern U.S., where each state serves as an observation. For an individual state, the values across each of the eight social determinants were assigned the mean value for each of the counties within that state. From here, we produced a two-dimensional map that shows the proximity of each state, having already calculated euclidean distances between SDH profiles.

Once we produced a mapping of the states, we then ran the same model, using each of the 1,422 counties in the southern US as our observations. We mapped the data, and estimated that they fell into roughly 3 regions in the two projected MDS coordinates. This observation was used to inform our choice of k in the

[8]This was not so at the state-wide level.

subsequent K-means cluster analysis on the projected 1,422 counties.

2.3 Regression Models

We considered three regression approaches to examine associations between HIV prevalence and demographic variables. First, we treated county-specific prevalence rates as a continuous outcome in a linear regression model. A logarithmic transformation was applied to the prevalence rates to accommodate the normality assumption. The second regression approach extends the linear regression model to include spatially-dependent random intercepts to account for potentially spatially-varying unmeasured confounders. This is accomplished through the use of the conditional autoregressive (CAR) model where spatial dependence is based on neighborhoods defined by the counties' geography. The third regression approach relaxes the normality assumption by modeling the cases as count data using Poisson regression, under a Bayesian hierarchical framework with CAR random effects.

2.3.1 Linear regression models

To examine factors associated with variation in HIV prevalence, linear regression models were first applied to the county level data.The general model is given by

$$\log(y_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon_i, \tag{1}$$

where y_i is the HIV prevalence rate for county i, x_i's are the socioeconomic and racial variables, and ε_i represents identically and independent residual error following $N(0, \sigma^2)$.

We first examined the pairwise relationship between HIV prevalence rates and each individual covariate.Next, we used multiple linear regression models to study the joint effect on HIV prevalence of all 8 predictors (percent of people aged above 18 without high school diploma, percent of poverty, median household income, urban indicator, unemployment rate, percent of the poverty, percent of non-hispanic black people, the percent of non-hispanic white people and percent of hispanic). We used the backward step-wise selection method to systematically remove predictors which are not significant based on whether removing a specific predictor gives a low AIC. We considered the entire southern United States as well as state specific models. Overall, 6 multiple linear regression models were selected. Model 1 represents the U.S. South regression analysis, and Model 2 to Model 6 are from the state-specific analysis. Model 2 includes all races, and Model 3 consists of the social-economic predictors which shows up the most in our state by state linear regression analysis. The remaining three models were formulated by considering one race and then including the two predictors which occurred most often with that race. All six models were applied to the entire U.S. South

(Table 3) and for each of the southern states individually excluding Delaware.

2.3.2 Conditional autoregressive models

Though the linear regression models are straightforward, the spatial potential dependence is ignored. The other two methods of conditional autoregressive model in this section and Hierarchical Bayesian conditional autoregressive model in the next section were employed to adjust the issue of spatial dependence.

One way to adjust for spatial dependece is to use conditional autoregressive model(CAR)[44]. The general model is given by

$$\log(y_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \theta_i + \varepsilon_i, \tag{2}$$

where an additional spatial component θ_i is included for county i. θ_i is defined conditionally as

$$\theta_i|\theta_{-i} \sim N\left(\frac{\sum_{j=1}^{n} w_{ij}\theta_j}{\sum_{j=1}^{n} w_{ij}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{ji}}\right), \tag{3}$$

$$\varepsilon_i \sim N(0, \sigma^2), \tag{4}$$

where $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$, and weight w_{ij} is 1 if county i, j share a common boundary or 0 otherwise. Therefore, spatial dependence of neighborhoods are defined based on county geography. Similar to the linear regression approach, we perform univariate analysis and the six selected multiple regression models shown in Table 2. Parameter estimation was carried out using maximum likelihood with the R package **spdep.**

2.3.3 Hierachical Bayesian conditional autoregressive models

Bayesian conditional autoregressive model [44] is another approach to account for the spatial heterogenity based on the Poisson model. Slightly different from the previous CAR model, the number of HIV diagonsis in a county is used instead of the HIV prevalence rate. We assumed that the number of observed HIV cases O_i of county i, conditional on the corresponding expected number of infected cases λ_i, is independently distributed as

$$O_i|\lambda_i \sim Poisson(\lambda_i), \tag{5}$$

and the Poisson mean λ_i is given by

$$\log(\lambda_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \log(P_i) + \theta_i + \varepsilon_i, \tag{6}$$

where P_i is the population at risk of county i, θ_i and ε_i are, respectively, spatial random effect and residual error, as described above. Under the Bayesian framework, we assumed priors of $\beta = (\beta_0, \cdots, \beta_p)$ follow an uninformative normal distribution $N(0, 1000I_p)$ and the priors of τ^2 and σ^2 follow $Inverse\ Gamma\ (0.001, 0.001)$. Parameter estimation was carried out using MCMC algorithm implemented in the R package **CARBayes**.

3 Results

A total of 1422 counties were included in this analysis. Summary statistics for the region and by state are given in Table 1. Note that over 40% of all persons living with HIV in the United States are in the south region.

We did not simply assign 0 for the suppressed case numbers for the following reason. We estimated that the overall prevalence of persons living with diagnosed HIV in the U.S. South to be 0.4%. This corresponds to a 0.90 probability of no HIV cases for counties with a population of 27 or less. Since the only county with population less than 100 with missing data had more than 27 people, we could not confidently assume there were no cases in the county. Overall, there were 168 counties (11.8%) with missing cases and rates. Among them, 130 counties had imputed values greater than 4 and were replaced with a value of 2 for number of cases.

3.1 Statistical Summary of Predictors and HIV rates

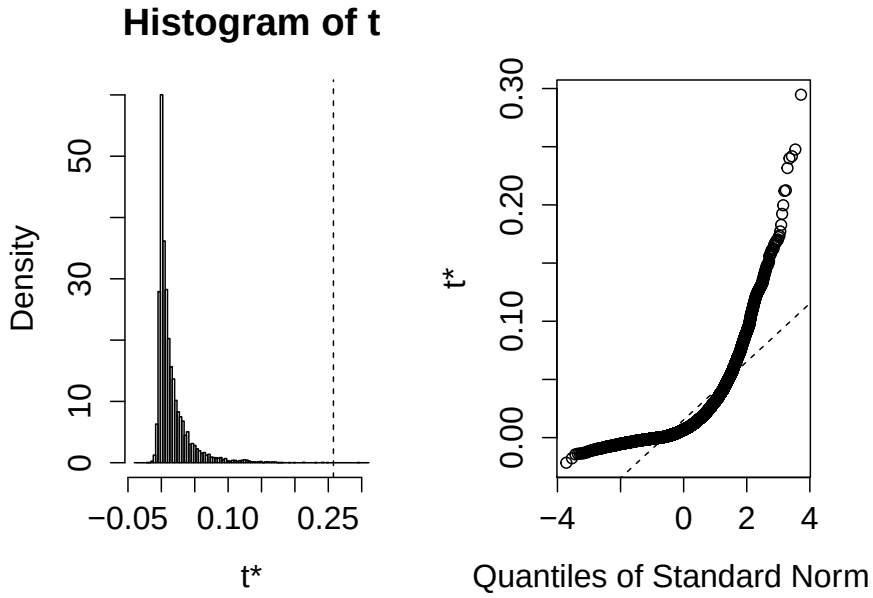
From Table 1, the mean HIV prevalence rate per 100,000 population for the U.S. South is 197.4, with Florida having the highest rate of 446.9, and Kentucky having the lowest rate of 71.7. There were clear spatial variation in HIV prevalence rates across the U.S. south The observed spatial trend was correlated with several racial and socioeconomic variables. For example, state-wide Florida had the second highest percent of Hispanics (following Texas) and the highest unemployment rate. The second highest HIV rate was 375.9 per 100,000 for South Carolina; which also had the second highest unemployment rate. Meanwhile, Kentucky has the least urban counties and second highest percent of people below the Federal Poverty Line. The second lowest HIV rate was 72.0 per 100,000 for West Virginia; which also has the lowest percent of NHB and Hispanics, with the highest percent of NHW.

For median income, Maryland has the highest across-county mean of $68,995 in relation to the U.S. South mean ($42,164), while Mississippi had the lowest of $34,473. Maryland had an HIV rate of 358.9, third highest for the U.S. South. Furthermore, Maryland has the lowest percent with less than a high school diploma, followed by Delaware; which is 100% urban and had the fourth highest HIV rate of 333.4 per 100,000. Alternatively, Mississippi had the highest percent of NHB in the U.S. south, second lowest percent of Hispanics,

and second highest percent of people with less than a high school diploma; along with the highest percent of people below the Federal Poverty Line, and a HIV rate of 266.0 per 100,000.

3.2 Global Measures of Spatial Clustering

We obtained the following distribution for Moran's index I, based on 10,000 non-parametric bootstrap replicas:



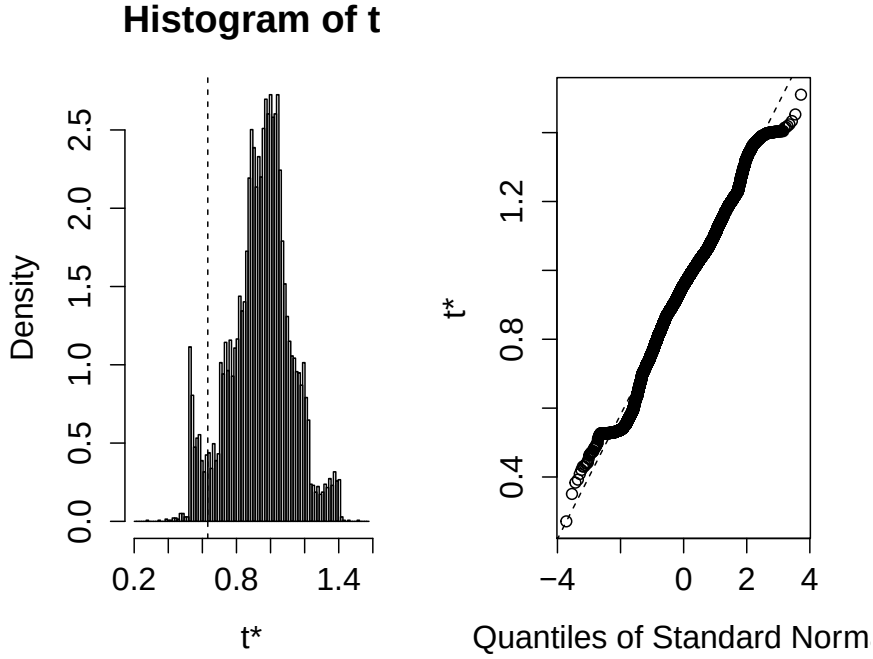
The value of the statistic based on SMR (standardized mortality ratio) is 0.258 with 95% confidence interval $(-0.0052, 0.08505)$.

Since $I = 0.258$ is positive, this suggests the presence of spatial clustering of HIV prevalence within the U.S. South. We note that the confidence interval includes zero suggesting some uncertainty in presence of the global spatial measure. Furthermore, as the magnitude of I is closer to zero than to one, we may not expect to see very high spatial correlation among regions with similar relative risks.

Geary's contiguity ratio c was found to have value

$$c = 0.632$$

with 95% confidence interval $(0.540, 1.317)$ based on the 10,000 bootstrap replicas of the Permutation Model :



Because the ratio c is less than one, there is evidence to suggest positive autocorrelation. Since the magnitude of the ratio is closer to 1 than to zero, this suggests that while there is some spatial clustering, the overall spatial clustering is not strong. This agrees with the estimated of Moran's I index for this dataset.

The results from Moran's I index and Geary's c ratio are in agreement, and this is not surprising based on their formulation. While Geary's c ratio is inversely related to Moran's I index, they are not identical. Moran's I is a measure of global spatial autocorrelation, while Geary's c is more sensitive to local spatial autocorrelation, by virtue of the terms involving $r_i - r_j$ in the computation. Besag-Newell clustering analysis was performed on the 1,422 counties in the southern U.S. to identify clusters of counties with high HIV prevalence. The 78 counties selected by the BNC analysis are highlighted in the map shown in Figure 6. The clusters capture isolated regions of Tennessee, Alabama, Georgia, Florida, Maryland, Virginia, and Delaware.

Table 11 contains a summary of the SDH values represented by the counties. Since the assumptions of normality are violated, we used the nonparametric Wilcoxon rank sum test to compare the SDH values in the clustered counties with the entire southern U.S. region. All of the SDH indicators in the cluster were significantly different from the regional indicators at the $\alpha = 0.05$ level, with the exception of the unemployment rates. Table 1 contains the mean and standard deviation for the regional SDH values. The map from a non-metric multidimensional scaling analysis is shown in Figure 2. The solution converged in 6 steps to a final $STRESS$ value of 2.32, indicating a good quality of fit. The Shepard plot for the solution exhibited strong

monotonicity and correlation.

States that are grouped close together should be considered most similar with respect to their S.D.H. averages. We see in the right-hand side that Delaware and Maryland are close together, away from the rest of the states. Texas is isolated at the top of the map. Among the rest of the states, Georgia and Tennessee are close together near the origin. No other state appears to neighbor another closer than Georgia and Tennessee.

We performed another non-metric MDS analysis using the counties as observations. The counties, plotted on the map according to state membership, are shown in the left-hand side of Figure 3. Based on this map, we decided to perform a K-Means clustering analysis, with $k = 3$. The plot on the right-hand side of Figure 3 shows the counties once again, colored and symbolized by cluster membership.

A table containing the cluster summaries is shown in Table 7. Assumptions of normality and homoskedasticity are violated by some of the samples, so we used Kruskal-Wallis to perform nonparametric tests for independence across each of the variables, grouping by cluster. All eight of the tests were significant at the $\alpha = 0.05$ level for the presence of at least one different sample. These results are shown in Table 8. To determine the samples that were different, we used Kruskal's multiple comparison test. The African American population in clusters 1 and 2 were not significantly different; all other comparisons tested as significantly different.

Finally, we ran a multiple linear regression within each cluster to test for the effect of the sampled S.D.Hs on the log-transformed observed HIV rates. The multiple R^2 values for clusters 1, 2, and 3 were 0.54, 0.52, and 0.50, respectively. The coefficients, as well as 95% confidence intervals, are shown in Table 9.

Figure 5 shows cluster membership for each county on a map of the southern U.S. We see a good deal of diversity for each of the three clusters. Cluster 1 contains areas in Texas, Mississippi, and Kentucky that are characteristically associated with areas of low income and low education. Similarly, cluster 3 contains counties with large metropolitan areas. Despite the geographic heterogeneity, as well as the differences between samples, the HIV rates per 100,000 are not significantly different according to a Kruskal-Wallis test. A boxplot for the rates is shown in Figure 4. We see that cluster 2 contains the counties showing the most extreme HIV rates. The mean HIV rate and standard deviation for each cluster are shown below in Table 10.

3.2.1 Multiple Regression Results

Table 3 reports the results for running the six multiple regression models (Table 2) with different sets of predictors using standard linear regression without accounting for potential spatial dependence. Overall, we found that percent NHB was positively associated with HIV prevalence, and this association is robust when adjusted for the proportions of other races and SES variables. SES variables including whether a county is urban, median income, and percent unemployment are the only predictors that had a positive association

with the prevalence of people living with HIV in all models in which they were included, noting that median income and unemployment were each included in only one model. Only the urban indicator and percent unemployment were statistically significant in all models in which they appeared (that is, the confidence interval did not include 1). These findings are consistent with those from the literature. On the other hand, percent NHW and percent with less than a high school education showed a negative association with HIV prevalence for all models in which they were included. This association is robust when adjusted for the other predictors. This finding is likely due to residual spatial confounding as previous studies have shown that HIV incidence is higher in regions with low educational attainment [13, 32, 33, 35].

In Table 4, we see that the CAR 1 method confirms percentage NHB, percentage below the poverty line, median income, whether a county is urban and unemployment as having a positive association with the rate of those living with HIV. This association is robust when adjusted for other predictors for percent NHB, percent below the poverty line, whether a county is urban and percent unemployment. Percent NHW and percentage of the population with less than a high school education are also negatively associated in the CAR 1 method. This association is robust when adjusted for other predictors for percent NHW.

Table 5 gives the estimates obtained from the CAR 2 model. Again, the results show that percent NHB, percentage of the population living below the poverty line and whether the county is urban had a robust positive association with HIV prevalence. Contrary to the linear regression and CAR 1 results, the risk ratio for unemployment given by CAR 2 is less than one after adjusting for percent Hispanic and percent living below the poverty line.

3.2.2 Univariate Analysis

Table 6 gives the estimated risk ratios and 95% confidence intervals for the univariate associations between HIV prevalence rates and the racial and SES variables obtained by the three regression approaches (standard linear regression, CAR 1 and CAR 2). The risk ratios for median income have been scaled by $10,000 and the risk ratios for the racial and SES variables correspond to a per unit increase in percentage.

Table 6 shows that, in most cases, the estimated risk ratios across the three methods are consistent in direction. For racial composition, all models indicate that percent NHB was positively associated with HIV prevalence with a risk ratio of while percent NHW was negatively associated with HIV prevalence. For SES variables, percent living below the poverty line, whether a county is urban, and percent unemployment were positively associated with HIV prevalence.

However, there are cases where the inclusion of spatial effects resulted in different risk ratios compared to the linear regression model. For median income, the standard linear regression model gave a risk ratio greater than one, indicating that increase in median income for a county would result in a positive increase in the

number of people living with HIV in that county. However, once the spatial effects were considered (CAR 1 and CAR 2), the risk ratio decreased to be below one, indicating that a rise in median income for a county was associated with a decline of cases of people living with HIV, which is what we would expect since HIV incidence has been shown to be higher in regions with low income [13, 32, 33, 35]. We see something similar with the percentage of people in a county with less than a high school education. Both the linear regression and CAR 1 models show that an increase in the percentage with less than a high school education would have a negative effect on HIV rate, while the CAR 2 model shows that percentage with less than a high school education has a positive and statistically significant effect on HIV prevalence. This is what we would expect since HIV incidence is higher in regions with low educational attainment [13, 32, 33, 35].

3.2.3 Cluster-Specific Regression Analysis

We performed a multiple linear regression analysis, testing the effect of the SDH indicators in the cluster against the log-transformed HIV rates ($F = 4.06$, $df = 8$ and 38, p-value $= 0.0014$). The results from this analysis are shown in Table 13. The 95% confidence interval for each of the coefficients contains 0.

We also used a Wilcoxon rank sum test to evaluate the observed HIV rates in the clustered counties for significance against the observed regional HIV rates ($W = 55595.5$, p-value < 0.001). This result indicates that the mean HIV rate in the clusters is different from the mean regional rate. Table 14 contains the mean and standard deviation for HIV rates in both the clustered counties and the entire region.

Finally, we ran a multiple linear regression within each cluster to test for the effect of the sampled S.D.Hs on the log-transformed observed HIV rates. The multiple R^2 values for clusters 1, 2, and 3 were 0.54, 0.52, and 0.50, respectively. The coefficients, as well as 95% confidence intervals, are shown in Table 9.

Figure 5 shows cluster membership for each county on a map of the southern U.S. We see a good deal of diversity for each of the three clusters. Cluster 1 contains areas in Texas, Mississippi, and Kentucky that are characteristically associated with areas of low income and low education. Similarly, cluster 3 contains counties with large metropolitan areas. Despite the geographic heterogeneity, as well as the differences between samples, the HIV rates per 100,000 are not significantly different according to a Kruskal-Wallis test. A boxplot for the rates is shown in Figure 4. We see that cluster 2 contains the counties showing the most extreme HIV rates. The mean HIV rate and standard deviation for each cluster are shown below in Table 10.

4 Discussion

4.1 Summary

The Besag-Newell cluster analysis results showed that counties in large metropolitan areas, including Memphis, Miami, Atlanta, D.C., and Baltimore had higher rates of persons living with HIV. Using these counties in the regression model, we determined that NHB remained the most important indicator for HIV prevalence rate in the southern United States. The regression analyses also indicated positive associations between county-level NHB proportion and HIV prevalence rates. Importantly, these associations remain even after adjusting for various socioeconomic variables and controlling for potential residual spatial confounding.

The K-Means analysis failed to partition the counties into any cluster whose HIV prevalence rate was significantly different from the regional HIV prevalence rate. However, it was successful at identifying NHB as the most important predictor for HIV prevalence. It is worth noting that the K-Means analysis partitioned the counties based solely on the proximity of their SDH scores, and completely independent of any geographic factors or HIV prevalence rates. Of course, a method like Besag-Newell is just the opposite, in that it only considers geography and disease prevalence. In the end, the results from Besag-Newell were more significant, which underscores the need for this type of spatial-based analysis when trying to understand patterns of disease spread on a map.

One of the problems we encountered while performing spatial clustering analysis with the Besag-Newell method was deciding on the window size, k. For a particular unit region, which in our case is a county, the method works by aggregating nearby counties until the number of reported cases within the accumulated counties exceeds this value of k. The particular issue we experienced relates to the subjectivity associated with choosing a value of k to use. This is a problem that has been discussed in previous literature [10, 36]. We included multiple values of k to detect recurring clusters as suggested by Costa and Assunçáo (2005) [10].

4.2 Limitations

For the analysis of our models we used the demographic variables restricted to the adult population of 18 and older, since the 13-18 year old population only constitutes a small percentage of each county. On average, 7% of the U.S. South population is between 15 to 19, with standard deviation 0.3% [39]. Since the percentages do not exceed 8% for any of the states in the U.S. South, it seems reasonable to compare the data despite this mismatch in age groups between the two data sources. Additionally, there is information from the NCHHSTP Atlas that 73.2% of the data for people living with diagnosed HIV in each county of the U.S. South for people ages 13-24 is suppressed [8]. Restricting our data analysis to only the U.S. South population 18 and above gives

a more accurate description of people living with diagnosed HIV. We restricted the percentage of persons with less than a high school education to 25 and above, in accordance with the 5-year ACS survey from 2008-2012.

When performing regression analysis, stepwise regression produces results that are easily misinterpreted. The elimination of a predictor by the method does not imply that the predictor is not the best predictor for the model. Rather, the elimination of a predictor at each step is conditional on which predictors are still left in the model. Despite the drawbacks of stepwise regression, we used it to gain some intuition as to which predictors would be best to include in candidate models without having to try all possible combinations of the predictors.

Results from the MCMC showed that the convergence of some chains are sensitive to the initial values even though a large portion of the samples is discarded through burn-in and thinning. One possible reason is that the prior distributions are non-informative, and so the starting values of chains from each run are different. The other possible reason is that some variables are highly correlated with the spatial adjustment of CAR assumption, so the simulation of the models with those variables seems hard to mix. One way to improve the convergence of chains is to take advantage of the information obtained from the linear regression or the conditional autoregressive model, which can be put into prior distributions. The other way is to increase the computational power to run longer simulation.

With respect to the global spatial clustering measures, according to [36], even though adjacency-based methods are easy to implement, they do not properly take into account differences in sampling variability of rates due to population heterogeneity." While there are some modifications to Moran's I index which can be explored, there are other distance-based measures such as those provided by Whittemore *et al.* (1987) and Tango (1995). Regions that are far away are weighted less in such implementations, so these measures may provide a different, and arguably more accurate picture of actual spatial clustering. While there are many problems associated with these types of global measures, taken together, they may help to paint a more accurate picture of the spatial clustering that occurs across the region. Furthermore, an indication of the existence of clustering, however rough, provides a good starting point for more localized analysis.

4.3 Future Work

To overcome the problem of choosing an appropriate k value in the Besag-Newell spatial clustering approach, Kuldorff and Nagarwalla (1995) [22] proposed fixing the fraction of the total at risk population instead of fixing the number of cases (k) in each window. With this approach each window now has the same number of people in it and sparse regions will necessarily be in larger windows. The above-mentioned methods ignore the geographical shapes of the regions, identifying each one of them with their geographical centroid or any other

relevant points within their boundaries. Cases and population are assumed to be located at these centroids. To take into account more spatial structure, some Bayesian methods have been proposed by Wakefield and Kim (2013) [43], and Lee and Mitchell (2013) [24, 25]. Another direction we would like to explore is to do state-specific spatial cluster analysis for multiple years. It would be interesting to see how the clusters identified at the state level compare to the clusters obtained at the regional level with more data for implementation.

Acknowledgements. This project was sponsored by the 2014 Industrial Math/Stat Modeling Workshop for Graduate Students, hosted by SAMSI at North Carolina State University. We are grateful to the organizers for this opportunity, especially Ilse Ipsen and Thomas Gehrmann.

References

[1] Adimora, Adaora A. and Victor J. Schoenbach. "Social Context, Sexual Networks, and Racial Disparities in Rates of Sexually Transmitted Infections." *The Journal of Infectious Diseases* 191 (2005): S115 – S122.

[2] Adimora, Adaora A., Victor J. Schoenbach, and Irene A. Doherty. "HIV and African Americans in the Southern United States: Sexual Networks and Social Context." *Sexually Transmitted Diseases* 33.7 (2006): S39 – S45.

[3] Adimora, Adaora A., Victor J. Schoenbach, Francis Martinson, Kathryn H. Donaldson, Tonya R. Stancil, and Robert E. Fullilove. "Concurrent Sexual Partnerships among African Americans in the Rural South." *Annals of Epidemiology* 14.3 (2004): 155 – 160.

[4] Aral, Sevgi O., Adaora A. Adimora, and Kevin A. Fenton. "Understand and responding to disparities in HIV and other sexually transmitted infections in African Americans." *The Lancet* 372 (2008): 337 – 340.

[5] Aral, Sevgi O., Ann O'Leary, and Charlene Baker. "Sexually Transmitted Infections and HIV in the Southern United States: An Overview." *Sexually Transmitted Diseases* 33.7 (2006): S1–S5.

[6] Berman, Stuart M., and Myron S. Cohen. "STD Treatment: How Can It Improve HIV Prevention in the South?" *Sexually Transmitted Diseases* 33.7 (2006): S50 – S57.

[7] Besag, Julian, and James Newell. "The Detection of Clusters in Rare Diseases." *Journal of the Royal Statistical Society, Series A* 154.1 (1991): 143 – 155.

[8] Centers for Disease Control and Prevention. (2010). NCHHSTP Atlas. Retrieved from http://gis.cdc.gov/GRASP/NCHHSTPAtlas/main.html.

[9] Cohen, Deborah A., Shin-Yi Wu, and Thomas A. Farley. "Structural Interventions to Prevent HIV/Sexually Transmitted Disease: Are They Cost-Effective for Women in the Southern United States?" *Sexually Transmitted Diseases* 33.7 (2006): S46 – S49.

[10] Costa, Marcelo Azevedo, and Renato Martins Assunção. "A fair comparison between the spatial scan and the Besag–Newell Disease clustering tests." *Environmental and Ecological Statistics* 12 (2005): 301 – 319.

[11] D'Aignaux, Jérôme Huillard, Simon N. Cousens, Nicole Delasnerie-Lauprêtre, Jean0Philippe Brandel, Dominique Salomon, Jean-Louis Laplanche, Jean-Jacques Hauw, and Annick Alpérovitch. "Analysis of the geographical distribution of sporadic Creutzfeldt-Jakob disease in France between 1002 and 1998." *Internation Journal of Epidemiology* 31 (2002): 490 – 495.

[12] Davids, Benem-Orom, Sonja Suzzete Hutchins, Camara P. Jones, and Jessie R. Hood. "Disparities in Life Expectancy Across US Counties Linked to County Social Factors, 2009 Community Health Status Indicators (CHSI)." *Journal of Racial and Ethnic Health Disparities* 1 (2014): 2 – 11.

[13] Farley, Thomas A. "Sexually Transmitted Diseases in the Southeastern United States: Location, Race, and Social Context." *Sexually Transmitted Diseases* 33.7 (2006): S58 – S64.

[14] Fleming, Patricia L., Amy Lansky, Lisa M. Lee, and Allyn K. Nakashima. "The Epidemiology of HIV/AIDS in Women in the Southern United States." *Sexually Transmitted Diseases* 33.7 (2006): S32 – S36.

[15] Geary, R.C. "The Contiguity Ratio and Statistical Mapping" *The Incorporated Statistician* 5 (3) (1954): 115 – 145

[16] Hall, H. Irene, Ruiguang Song, Philip Rhodes, Joseph Prejean, Qian An, Lisa M. Lee, John Karon, Ron Brookmeyer, Edward H. Kaplan, Matthew T. McKenna, Robert S. Janssen. "Estimation of HIV Incidence in the United States." *Journal of the American Medical Association* 300.5 (2008): 520 – 529.

[17] Harrison, Kathleen McDavid, Qiang Ling, Ruiguang Song, and H. Irene Hall. "County-Level Socioeconomic Status and Survival After HIV Diagnosis, United States." *Annals of Epidemiology* 18.12 (2008): 919 – 927.

[18] Heckman, Timothy G., Anton M. Somlai, Seth C. Kalichman, Stephen L. Franzoi, and Jeffrey A. Kelly. "Psychosocial Differences Between Urban and Rural People Living With HIV/AIDS." *The Journal of Rural Health* 14.2 (1998): 138 – 145.

[19] Heckman, T. G., A. M. Somlai, J. Peters, J. Walker, L. Otto-Salaj, C. A. Galdabini, and J. A. Kelly. "Barriers to care among persons living with HIV/AIDS in urban and rural areas." *AIDS Care: Psychological and Socio-medical Aspects of AIDS/HIV* 10.3 (1998): 365 – 375.

[20] Jia, Haomiao, David G. Moriarty, and Norma Kanarek. "County-Level Social Environment Determinants of Health-Related Quality of Life Among US Adults: A Multilevel Analysis." *Journal of Community Healthy* 34.5 (2009): 430 – 439.

[21] Jones, Camara Phyllis. " 'Race,' Racism, and the Practice of Epidemiology." *American Journal of Epidemiology* 154.4 (2001): 299 – 304.

[22] Kuldorff M., Nagarwalla N. "Spatial disease clusters: Detection and Inference." *Statistics in Medicine*, 14 (1995): 799 – 810

[23] Lattin, James, J. Douglas Carroll, and Paul E. Green. *Analyzing Multivariate Data*. Belmont, CA: Brooks/Cole, Cengage Learning, 2003. Print.

[24] Lee, Dunca. "CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors." *Journal of Statistical Software* 55.13 (2013)

[25] Lee, D. and R. Mitchell. "Boundary detection in disease mapping studies." *Biostatistics*, 13 (2012): 415 – 426

[26] Lichtenstein, Bronwen. "Stigma as a barrier to treatment of sexually transmitted infection in the American deep south: issues of race, geneder and poverty." *Social Science & Medicine* 57 (2003): 2435 – 2445.

[27] Marmot, Michael. "Social determinants of health inequalities." *The Lancet* 365 (2005): 1099 – 1104.

[28] Moran, P.A.P. "Notes on Continuous Stochastic Phenomena." *Biometrika* 37 (1) (1950): 17 – 23.

[29] Murray, Christopher J. L., Sandeep C. Kulkarni, Catherine Michaud, Niels Tomijima, Maria T. Bulzacchelli, Terrell J. Iandiorio, and Majid Ezzati. "Eight Americas: Investigating Mortality Disparities across Races, Counties, and Race-Counties in the United States." *PLoS Medicine* 3.9 (2006): 1513 – 1524.

[30] O'Leary, Ann, Sherry D. Broadwell, Peikang Yao, and Deborah Hasin. "Major Depression, Alcohol and Drug Use Disorders Do Not Appear to Account for the Sexually Transmitted Disease and HIV Epidemic in the Southern United States." *Sexually Transmitted Diseases* 33.7: S70 – S77.

[31] Probst, Janice C., James N. Laditka, and Sarah B. Laditka. "Association between community health center and rural health clinic presence and county-level hospitalization rates for ambulatory care sensitive conditions: an analysis across eight US states." *BMC Health Services Research* 134.9 (2009): 11 pages.

[32] Reif, Susan, Kristin Lowe Geonnotti, and Kathryn Whetten. "HIV Infection and AIDS in the Deep South." *American Journal of Public Health* 96.6 (2006): 970 – 973.

[33] Reif, S., C. E. Golin, and S. R. Smith. "Barriers to accessing HIV/AIDS care in North Carolina: Rural and urban differences." *AIDS Care* 17.5 (2005): 558 – 565.

[34] Smith, David Barton. "Racial Disparities in Care: The Concealed Legacy of a Divided System." *Sexually Transmitted Diseases* 33.7 (2006): S65 – S69.

[35] Song, Ruiguang, H. Irene Hall, Kathleen McDavid Harrison, Tanya Telfair Sharpe, Lillian S. Lin, and Hazel D. Dean. "Identifying the Impact of Social Determinants of Health on Disease Rates Using Correlation Analysis of Area-Based Summary Information." *Public Health Reports* 126 (2011): 70–80.

[36] Tango, Toshiro. *Statistical Methods for Disease Clustering.* Springer Science & Business Media, 2010. Google eBook.

[37] Thomas, James C. "From Slavery to Incarceration: Social Forces Affecting the Epidemiology of Sexually Transmitted Diseases in the Rural South." *Sexually Transmitted Diseases* 33.7 (2006): S6 – S10.

[38] Thomas, Kathleen C., Alan R. Ellis, Thomas R. Konrad, Charles E. Holzer, and Joseph P. Morrissey. "County-Level Estimates of Mental Health Professional Shortage in the United States." *Psychiatric Services* 60.10 (2009): 1323 – 1328.

[39] U.S. Census Bureau; American Community Survey, 2008-2012 American Community Survey 5-Year Estimates, American FactFinder; <http://factfinder2.census.gov>; (15 July 2014).

[40] U.S. Census Bureau. (7 March 2013). Geography: Urban and Rural Classification. Retrieved from <https://www.census.gov/geo/reference/urban-rural.html>

[41] U.S. Census Bureau. Geography: Census Regions and Divisions of the United States. Retrieved from <https://www.census.gov/geo/maps- data/maps/pdfs/reference/us_regdiv.pdf>

[42] U.S. Census Bureau. American Community Survey Office: When to use1-year, 3-year, or 5-hear estimates. (27 June 2014) Retrieved from <http://www.census.gov/acs/www/guidance_for_data_users/estimates/>

[43] Wakefield J. and Kim A.Y. "A Bayesian model for cluster detection." *Biostatistics*, 14 (2013): 752 – 765

[44] Waller, Lance A., and Carol A. Gotway. *Applied Spatial Statistics for Public Health Data.* Hoboken, NJ: Wiley-Interscience, 2004.

[45] Wheeler, Darrell P. "Exploring HIV Prevention Needs for Nongay-Identified Black and African American Men Who Have Sex With Men: A Qualitative Exploration." *Sexually Transmitted Diseases* 33.7 (2006): S11 – S16.

[46] Will, Julie C., Isaac A. Nwaise, Linda Schieb, and Yuna Zhong. "Geographic and Racial Patterns of Preventable Hospitalizations for Hypertension: Medicare Beneficiaries, 2004 – 2009." *Public Health Reports* 129 (2014): 8 – 18.

5 Tables

		U.S. South	AL	AR	DE	FL	GA	KY	LA	MD	MS	NC	OK	SC	TN	TX	VA	WV
HIV Rates	*Mean*	197.4	204.7	131.8	333.4	446.9	245.6	71.7	293.8	358.9	266.0	230.2	84.1	375.9	118.3	142.8	231.0	72.0
	S.D.	237.2	143.6	116.7	100.3	439.2	238.6	55.4	266.6	473.5	176.7	146.5	54.4	185.3	126.1	252.5	235.8	69.9
	IQR	173.3	162.5	84.55	116.1	291.1	197.6	61.0	221.7	255.8	206.7	237.1	40.5	286.8	60.4	98.3	234.2	52.9
% NHB	*Mean*	18.7	26.2	15.4	20.9	15.2	30.2	7.7	31.8	29.0	37.0	21.2	7.1	27.7	16.6	11.5	19.1	3.1
	S.D.	18.0	22.4	17.8	5.1	9.4	17.5	4.2	14.4	16.4	20.7	16.3	3.5	16.5	10.4	6.8	16.7	2.5
	IQR	24.7	31.0	27.7	5.4	8.9	25.3	4.3	21.6	14.8	31.6	27.7	3.1	22.0	6.3	7.8	24.3	2.7
% NHW	*Mean*	60.1	67.0	74.5	65.3	57.8	55.8	86.3	60.3	54.7	58.0	65.2	68.7	64.1	75.6	45.3	64.8	93.0
	S.D.	20.3	20.7	17.1	5.9	14.6	17.0	5.9	13.5	19.4	19.8	17.5	9.3	15.1	11.5	21.2	18.0	3.8
	IQR	29.5	27.1	27.5	6.9	17.0	20.9	6.7	14.8	19.4	29.3	29.8	12.7	22.5	8.6	29.8	27.6	3.8
% Hisp.	*Mean*	15.9	3.8	6.4	8.1	22.5	8.8	3.0	4.3	8.2	2.6	8.3	8.8	5.0	4.5	37.6	7.9	1.2
	S.D.	15.3	2.9	4.9	1.3	11.9	5.2	1.6	2.0	3.6	1.9	3.8	6.6	3.0	2.3	23.1	5.2	1.0
	IQR	7.8	2.3	3.1	1.4	12.1	5.0	1.7	2.2	2.2	1.7	5.0	5.6	2.9	2.1	30.4	3.9	0.6
% Less than HS	*Mean*	16.0	17.3	16.7	12.3	14.2	15.6	17.5	17.8	11.6	18.9	15.4	13.8	15.9	16.1	19.3	13.1	16.5
	S.D.	6.7	4.9	4.7	1.9	6.7	6.2	7.4	5.7	4.0	6.0	5.0	4.1	5.1	5.2	8.2	6.6	6.0
	IQR	9.0	6.5	6.9	2.2	9.3	7.8	11.7	9.0	5.7	7.6	7.6	5.3	7.4	7.5	8.8	10.0	7.4
% Below FPL	*Mean*	14.2	15.6	16.0	9.9	13.8	15.2	16.4	16.1	8.6	19.0	14.6	14.3	15.2	15.0	14.6	10.2	15.9
	S.D.	5.8	5.3	3.8	0.3	4.5	5.8	6.3	5.3	4.3	6.1	4.1	4.1	4.6	4.2	5.4	6.7	4.1
	IQR	6.9	6.2	4.7	0.4	6.6	7.6	9.1	5.8	5.1	6.9	5.3	5.0	5.6	4.9	5.3	10.0	4.8
Med. income	*Mean*	42164	37812	35855	57716	43875	40384	37922	41977	68995	34473	41673	42166	39229	39180	44957	52561	37781
	S.D.	12244	8310	5943	5076	7372	11498	10109	9229	19758	7773	7596	7250	7855	8862	10344	19424	6755
	IQR	11909	9966	7111	5989	11289	11244	13679	11359	31696	7351	9936	9710	9902	9056	11227	23900	6423
% Urban	*Mean*	29.9	40.3	18.7	100.0	61.2	25.8	13.3	35.9	70.8	17.1	53.0	19.5	52.2	30.5	24.4	26.1	20.0
Unemp. rate	*Mean*	9.4	10.4	8.7	8.4	11.4	10.7	9.6	8.5	7.9	10.8	10.6	6.8	11.2	9.9	7.8	7.1	7.9
	S.D.	3.6	4.0	2.8	0.5	2.7	3.2	2.9	3.3	2.1	4.2	2.7	2.5	2.9	2.6	2.8	2.9	2.4
	IQR	4.3	5.2	4.0	0.6	3.5	4.3	2.9	4.0	2.7	4.6	3.1	3.6	3.7	3.9	3.3	3.8	2.8

Table 1: Displays the rate (per 100,000) of people living with diagnosed HIV in 2010 and different predictors considered for the whole U.S. South and each of the 16 states in the U.S. South from the 5-year ACS for 2008-2012. For each predictor there is the mean, standard deviation (S.D.), and IQR (inner quartile range); except for % Urban, which is the percent of urban counties.

Model	Response	Predictors (Variables)
1	HIV incidence rate (log)	WHS, Income, Urban, NHB(%), NHW(%)
2	HIV incidence rate (log)	NHB(%), NHW(%) , Hispanic(%)
3	HIV incidence rate (log)	WHS(%), Poverty, Urban Indicator
4	HIV incidence rate (log)	WHS(%), Urban Indicator, NHB(%)
5	HIV incidence rate (log)	WHS(%), Poverty(%) , NHW(%)
6	HIV incidence rate (log)	Poverty(%), Unemployment(%), Hispanic(%)

Table 2: Multiple linear regression models.

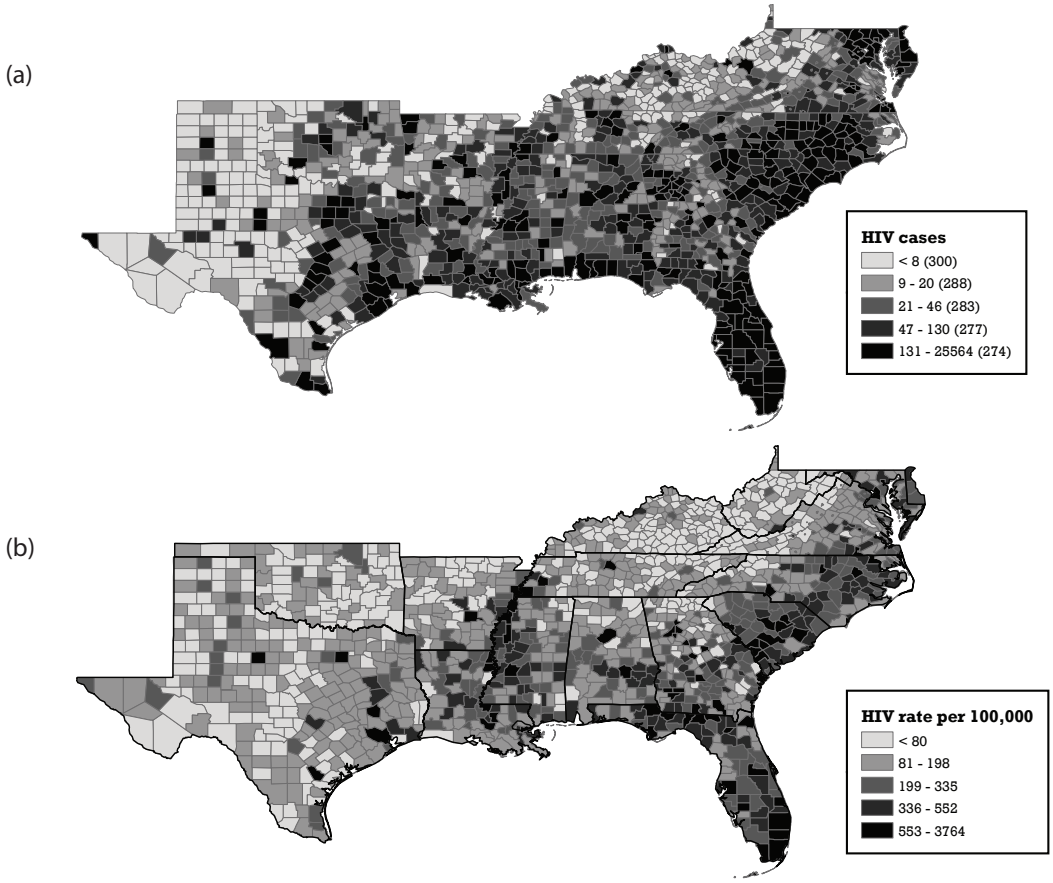


Figure 1: Maps showing HIV cases (a) and rates (b) in southern U.S. counties.

	1	2	3	4	5	6
NHB	1.026	1.016		1.033		
	(1.023, 1.028)	(1.008, 1.025)		(1.031, 1.035)		
NHW	0.989	0.981			0.973	
	(0.987, 0.992)	(0.973, 0.989)			(0.971, 0.975)	
Hisp.		0.991				1.003
		(0.983, 0.999)				(1.000, 1.006)
Less than HS	0.988		0.986	0.994	0.955	
	(0.980, 0.995)		(0.977, 0.995)	(0.988, 1.000)	(0.948, 0.962)	
Below FPL			1.041		1.024	0.992
			(1.031,1.052)		(1.016, 1.033)	(0.983, 1.002)
Med. income (per 10,000)	1.036					
	(0.996, 1.078)					
Urb. ind.	1.443		1.776	1.596		
	(1.325,1.571)		(1.592,1.982)	(1.466,1.737)		
Unemp.						1.078
						(1.062,1.095)

Table 3: Table giving the estimated risk ratios (top) and 95% confidence intervals (bottom) for the associations between HIV prevalence rates (dependent variable) and the racial and SES variables obtained by linear regression. The estimates are given for each of the six multiple regression models (Table 2).

	1	2	3	4	5	6
NHB	1.025	1.011		1.034		
	(1.021, 1.029)	(1.000, 1.022)		(1.031, 1.037)		
NHW	0.989	0.977			0.973	
	(0.985, 0.992)	(0.966, 0.987)			(0.971, 0.976)	
Hisp.		0.985				1.002
		(0.9740.996)				(0.998, 1.007)
Less than HS	0.986		0.992	0.995	0.967	
	(0.978, 0.995)		(0.983, 1.000)	(0.988, 1.001)	(0.967, 0.974)	
Below FPL			1.039		1.017	1.019
			(1.029, 1.049)		(1.008, 1.026)	(1.009,1.029)
Med. income (per 10,000)	1.006					
	(0.961, 1.053)					
Urb. Ind.	1.320		1.509	1.415		
	(1.208,1.444)		(1.368,1.666)	(1.297,1.544)		
Unemp.						1.024
						(1.008,1.039)

Table 4: Table giving the estimated risk ratios (top) and 95% confidence intervals (bottom) for the associations between HIV prevalence rates (dependent variable) and the racial and SES variables obtained by CAR 1. The estimates are given for each of the six multiple regression models (Table 2).

	1	2	3	4	5	6
NHB	1.016	MDC		1.030		
	$(1.009, 1.023)$	(MDC, MDC)		$(1.026, 1.034)$		
NHW	0.985	MDC			0.972	
	$(0.978, 0.990)$	(MDC, MDC)			$(1.967, 0.979)$	
Hisp.		MDC				0.976
		(MDC, MDC)				$(0.973, 1.980)$
Less than HS	0.990		MDC	1.004	0.978	
	$(0.981, 0.999)$		(MDC, MDC)	$(0.997, 1.012)$	$(0.966, 0.990)$	
Below FPL			MDC		1.010	1.115
			(MDC, MDC)		$(1.000, 1.019)$	(1.102,1.129)
Med. income (per 10,000)	0.926					
	$(0.885, 0.993)$					
Urb. ind.	1.267		MDC	1.402		
	(1.096,1.472)		(MDC, MDC)	(1.245,1.560)		
Unemp.						0.972
						$(0.946, 0.997)$

Table 5: Table giving the estimated risk ratios (top) and 95% confidence intervals (bottom) for the associations between HIV prevalence rates (dependent variable) and the racial and SES variables obtained by CAR 2. The estimates are given for each of the six multiple regression models shown in Table 2. Note that MDC indicates the model did not converge.

	LR	CAR1	CAR2
NHB	1.033	1.034	1.014
	(1.031, 1.035)	(1.031, 1.037)	(1.009, 1.022)
NHW	0.975	0.976	0.974
	(0.973, 0.977)	(0.973, 0.978)	(0.969, 0.979)
Hisp.	0.999	1.004	0.977
	(0.996, 1.002)	(0.999, 1.008)	(0.973, 0.981)
Less than HS	0.991	0.996	1.064
	(0.984, 0.998)	(0.989, 1.003)	(1.042,1.084)
Below FPL	1.018	1.026	1.120
	(1.009,1.026)	(1.018, 1.035)	(1.103, 1.129)
Med. income (per 10,000)	1.039	0.953	0.631
	(0.999, 1.081)	(0.913, 0.994)	(0.564, 0.739)
Urb. ind.	1.676	1.450	1.304
	(1.514, 1.855)	(1.323, 1.590)	(1.094, 1.642)
Unemp.	1.07	1.04	1.037
	(1.05, 1.08)	(1.02, 1.05)	(1.021, 1.049)

Table 6: Table giving the estimated risk ratio (top) and 95% confidence interval (bottom) for the univariate associations between HIV prevalence rates (dependent variable) and the eight racial and SES variables obtained from linear regression (LR), CAR1, and CAR2.

	HS or Less	Below FPL	Med. Income	Urbanicity	Unemployment	NHB	NHW	Hisp
Cluster 1								
Mean	29.65	23.94	10.34	0.052	12.76	0.26	0.57	0.15
S.D.	5.27	5.20	0.17	0.22	4.48	0.26	0.28	0.26
Cluster 2								
Mean	21.48	17.75	10.54	0.22	10.02	0.17	0.71	0.085
S.D.	3.40	3.27	0.14	0.41	3.02	0.17	0.18	0.13
Cluster 3								
Mean	13.89	11.51	10.84	0.52	7.74	0.12	0.75	0.084
S.D.	3.36	3.41	0.22	0.50	2.61	0.12	0.14	0.086

Table 7: Cluster summaries for the K-Means clustering solution using projected MDS solution.

	Less than HS	Below FPL	Med. Income	Urban ind.	Unemployed	NHB	NHW	Hisp
K.W. χ^2	1026.51	902.72	822.26	211.79	311.86	27.42	74.21	53.75
p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Table 8: Kruskal-Wallis χ^2 statistics and corresponding p-values for the significance tests of each of the mean SDH values between clusters.

Variable	β	*S.E.*	2.5%	97.5%
Cluster 1				
Less than HS	0.99	0.01	(0.96,	1.01)
Below FPL	1.02	0.01	(0.99,	1.05)
Med. Income	2.7	0.47	(1.07,	6.85)
Urban ind.	1.02	0.23	(0.65,	1.6)
Unemployed	1.02	0.01	(0.99,	1.05)
NHB	0.91	1.64	(0.04,	22.69)
NHW	0.04	1.64	(0,	1.05)
Hisp.	0.13	1.65	(0,	3.17)
$F = 36.03$, $df = 8$ and 240, p-value < 0.001, $R^2 = 0.55$.				
Cluster 2				
Less than HS	1	0.01	(0.99,	1.02)
Below FPL	1.03	0.01	(1.01,	1.05)
Med. Income	3.57	0.27	(2.09,	6.08)
Urban ind.	1.36	0.07	(1.2,	1.54)
Unemployed	1.01	0.01	(0.99,	1.02)
NHB	24.33	0.54	(8.41,	70.33)
NHW	0.58	0.54	(0.2,	1.67)
Hisp.	1.25	0.59	(0.39,	3.96)
$F = 87.23$, $df = 8$ and 648, p-value < 0.001, $R^2 = 0.52$.				
Cluster 3				
Less than HS	0.98	0.01	(0.96,	1)
Below FPL	1.01	0.01	(0.99,	1.04)
Med. Income	1.24	0.24	(0.79,	1.97)
Urban ind.	1.35	0.07	(1.18,	1.53)
Unemployed	1.02	0.01	(1,	1.05)
NHB	34.04	0.67	(9.09,	127.42)
NHW	0.61	0.63	(0.18,	2.09)
Hisp.	2.66	0.72	(0.65,	10.83)
$F = 63.1$, $df = 8$ and 507, p-value < 0.001, $R^2 = 0.50$.				

Table 9: Coefficients, standard errors, and 95% confidence intervals for multiple linear regression analysis within each cluster.

	Cluster 1	Cluster 2	Cluster 3
Mean	224	200	182
S.D.	207	282	180
W	165178	474346	371518
p-value	0.091	0.57	0.67

Table 10: The mean and standard deviation for HIV prevalence rate in each cluster, and the results from a Wilcoxon rank sum test for significance against the regional HIV prevalence rate. The three clusters are not significantly different according to a Kruskal-Wallis rank sum test ($K.W.$ $\chi^2 = 4.07$, $df = 2$, p-value $= 0.13$).

	HS or Less	Below FPL	Med. Income	Urbanicity	Unemployed	NHB	NHW	Hisp
Mean	15.64	12.8	10.87	0.68	10.54	0.24	0.61	0.11
S.D.	6.1	4.8	0.26	0.47	3.26	0.16	0.18	0.12

Table 11: Mean and standard deviation for each of the SDHs in the counties included in the BNC solution.

	β	*S.E.*	2.5%	97.5%
HS or Less	0.04	0.02	(1,	1.08)
Below FPL	0.08	0.02	(1.04,	1.13)
Med. Income	-1.28	0.43	(0.12,	0.65)
Urbanicity	0.07	0.26	(0.65,	1.78)
Unemployed	0.06	0.04	(0.99,	1.14)
NHB	2.23	0.68	(2.43,	35.5)
NHW	-2.61	0.53	(0.03,	0.21)
Hisp	1.86	0.95	(1,	41.23)

Table 12: Simple linear regression coefficients for each of the SDH values as an explanatory variable for the log-transformed HIV rates in the clustered counties.

	β	*S.E.*	2.5%	97.5%
Less than HS	1	0.03	(0.93,	1.06)
Med. income	0.44	0.63	(0.13,	1.52)
Urban ind.	0.93	0.31	(0.51,	1.69)
NHB	1.07	0.97	(0.16,	7.13)
NHW	0.1	0.95	(0.02,	0.63)
F-statistic: 5.65, $df = 5$ and 41, p-value $= 0.00047$				

Table 13: Multiple linear regression analysis coefficients for the SDH indicators ($F = 5.65$, $df = 5$ and 41, p-value $= 0.00047$, $R^2 = 0.41$).

	Cluster	U.S. South
Mean	596.95	197.37
S.D.	610.78	237.26

Table 14: Mean and standard deviation for the HIV rates in the clustered counties and the U.S. south region ($W = 55595.5$, p-value < 0.001).

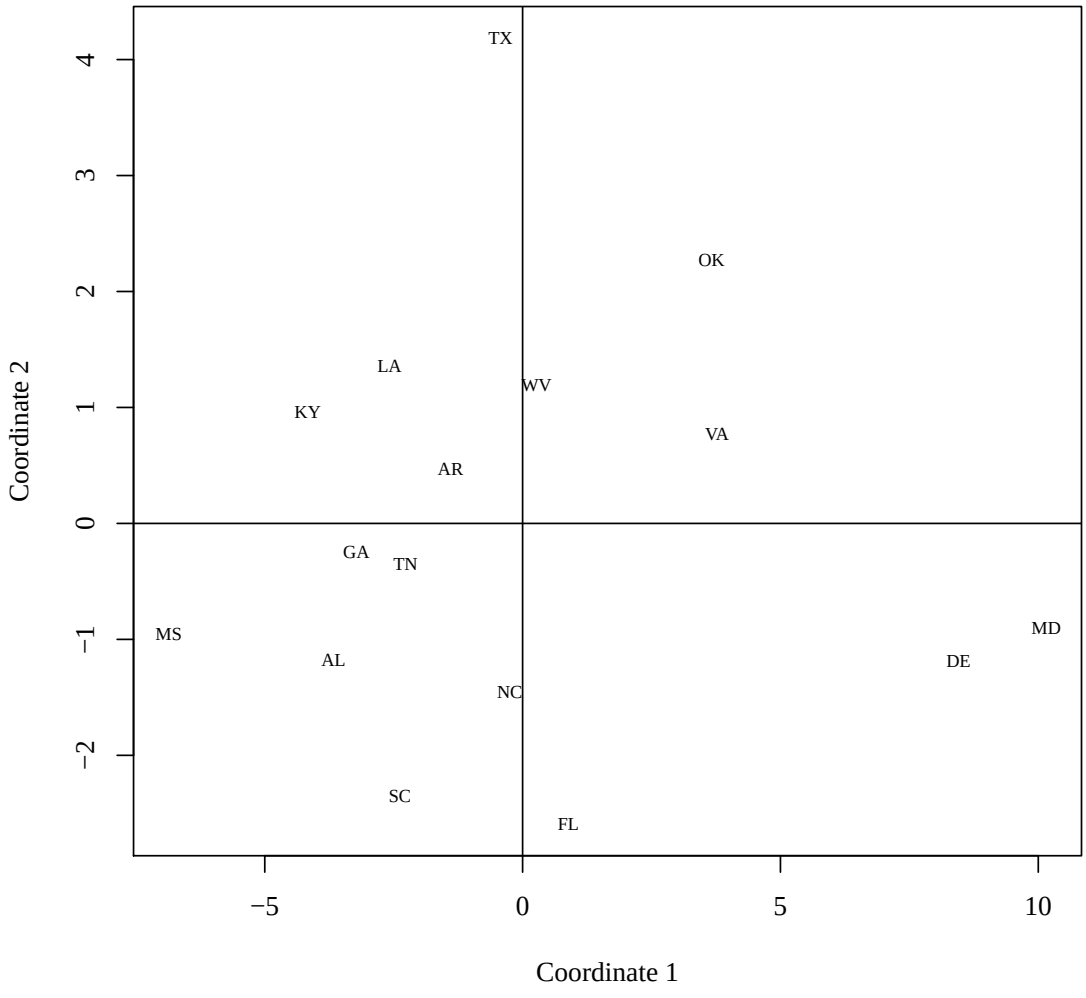


Figure 2: Multidimensional scaling solution, with the states used as observations.

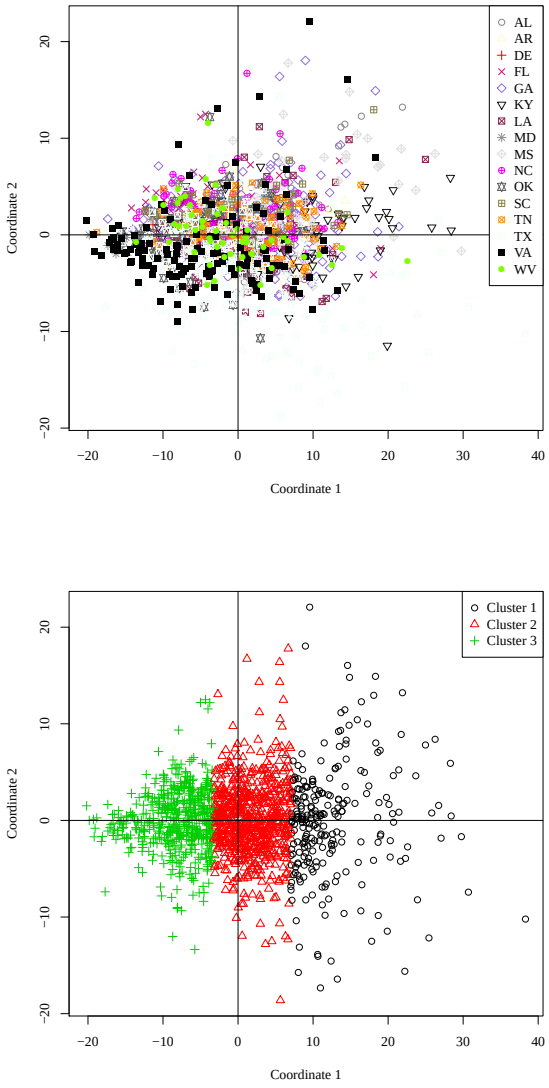


Figure 3: MDS solution with the counties as observations. The left-hand side shows membership according to state, while the right-hand side shows cluster membership after a K-Means clustering analysis, $k = 3$.

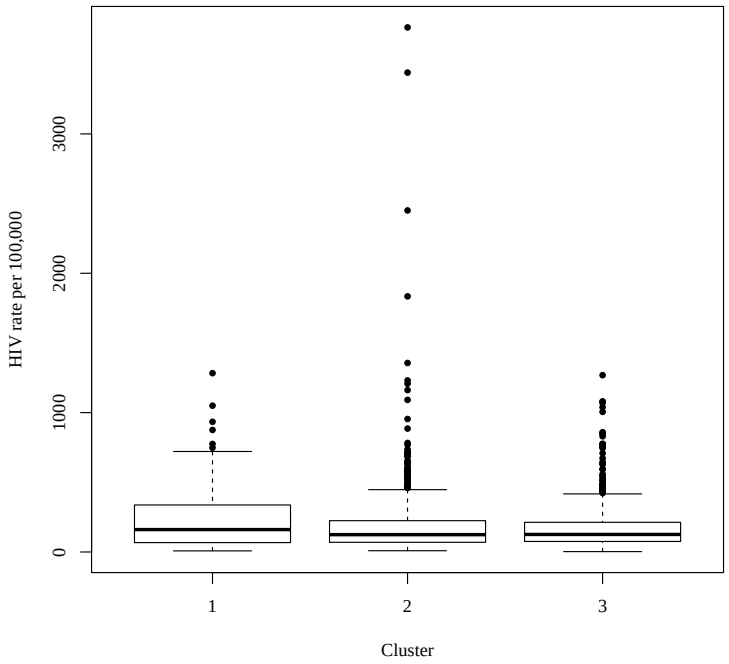


Figure 4: Boxplot of the HIV rate per 100,000 for each of the three clusters.

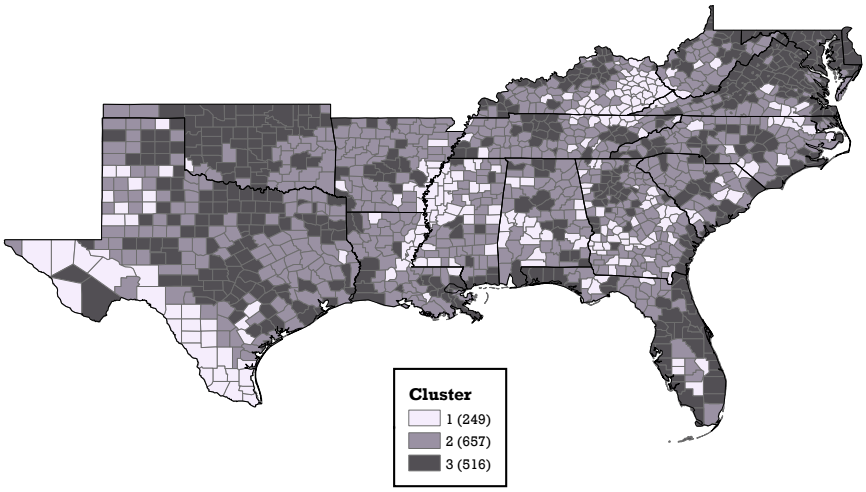


Figure 5: Map of the southern U.S., with counties colors according to cluster membership.

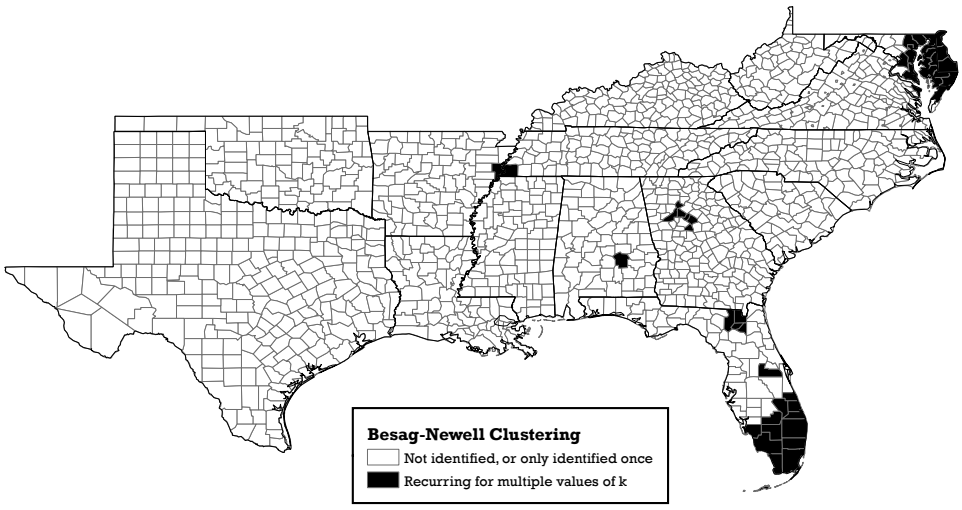


Figure 6: Map showing the counties identified by a Besag-Newell clustering analysis as having a high HIV prevalence.

Allergies, Asthma and Exposures in the Homes of the U.S. Population

Alexej Gossmann[1], Anastasia Wilson[2], Hongjuan Zhou[3], Nancy Hernandez Ceron[4], Tamra Heberling[5]
Yuanzhi Li[6]

Industry Mentors: Agustin Calatroni[7], Herman Mitchell[7], Russ Helms[7]
Faculty Mentors: Sanvesh Srivastava[8]

Abstract

While only 9% of the population suffers from asthma, the cost to patients and the health care system is significant. Currently, more needs to be discovered about the cause of asthma, and preventative measures. For this reason a statistical analysis of the National Health and Nutrition Examination Study (NHANES) 2005-2006 survey was performed to develop a prediction model for asthma based on allergies and exposures. While there are many methods available we focused primarily on logistic regression and random forests classification for predicting asthma from exposures and allergies, and logistic regression, LASSO regression and random forest for predicting exposures based on home environmental factors. For this type of data set where there can be significant collinearity between the variables, we found that random forests had the highest accuracy rate for prediction in both cases. Ethnicity and exposure to pets such as cats and dogs as well as exposure to dust mites contributed greatly to asthma. We investigated the relationship between home factors and exposure to cockroaches since an allergy to cockroaches has been shown to directly correlate with asthma. We found ethnicity, income, education level, and various home factors such as the size of the home, the number of individuals living in the home, and how long ago the home was built, all had significant effects on cockroach exposures.

1 Introduction

Asthma is a disease that effects the quality of life in patients, and places considerable economic burden on the United States health care system. In 1998, asthma in the U.S. cost approximately 12.7 billion dollars annually [15]. The prevalence of asthma has also increased dramatically in the past few decades jumping 171% between 1980 and 2010 [13]. Although the increase in asthma slowed somewhat from 2001 to 2010, the number of people suffering from asthma still increased an average of 2.9% each year [12], and it is now at its highest level [2].

Increases in allergic sensitivities has been equally substantial in recent decades [7]. Now 45% of the U.S. population is allergic to at least one allergen [14] with dust mites being one of the most common airborne allergens worldwide [8]. Food allergies among children have increased significantly in the last decade [4], which is especially significant since children with a food allergy are two to four times as likely to have other allergic sensitivities and asthma than children who do not have a food allergy [3]. The specific cause of asthma is not known, but it is generally accepted to be a combination of environmental exposures and allergic sensitivities [5, 10]. Atopic patients exposed to house dust mite allergens can develop allergic asthma and other inflammatory diseases [8]. Environmental exposures in allergic individuals can cause an asthmatic reaction in persons suffering from asthma. Though not every asthmatic individual is sensitive to allergens and not every allergic individual develops asthma, there is a strong correlation between the two.

The Center for Disease Control and Prevention (CDC) recently conducted a nationwide survey that has allowed further study of the relationship between asthma, allergies and exposures. The National Health and Nutrition Examination Survey (NHANES) 2005-2006 consisted of an interview containing demographic,

[1]Department of Mathematics, Tulane University
[2]Department of Mathematical Sciences, Clemson University
[3]Department of Mathematics, University of Kansas
[4]Department of Mathematics, Purdue University
[5]Department of Mathematical Sciences, Montana State University
[6]Department of Mathematics and Statistics, Utah State University
[7]Rho Inc.
[8]SAMSI

socioeconomic, dietary, and health- and home-related questions as well as a medical examination and home dust collection. The medical examination included a blood sample which was analyzed in an on-site laboratory, and the home dust collection was later analyzed in a laboratory for common allergen exposures. NHANES 2005-2006 was the first study of its kind to compile information on asthma, allergies and possible allergen exposures in the home. Salo et. al. previously analyzed the demographic, socioeconomic, asthma and allergen sensitization data [14]. However, the home exposure data was released only recently, and as a result, we are only now able to analyze the data to determine the link(s) between allergen sensitization, asthma, and home exposure. In fact, this is the first time in history that we have nationwide data to allow for this type of analysis.

This paper is organized in the following way. In section 2, we describe in detail the problem we wish to model and the data set used to formulate the model. In section 3, we will detail the logistic regression model and the random forest model that we used. The results and a comparison of the two models can be found in section 4. The paper will close, in section 5, with a summary.

2 The Problem

Our main objective was to build a predictive model for asthma based on home exposures and allergy sensitivities. Identifying the home exposures and allergies with a high correlation to asthma would allow healthcare professionals to provide targeted preventative measures for families with asthma across the United States. In addition, we wanted to predict the home exposures based on questionnaire data in order to identify important factors that lead to home exposure.

The data used to develop the models was obtained from NHANES 2005-2006. NHANES 2005-2006 is unique in that it included information on home exposures in addition to the general health questions and laboratory tests involved in previous NHANES studies. The process used to obtain the demographic, socioeconomic, dietary, and medical data was summarized in Salo et. al. [14]. Dust samples were obtained via home visits where two CDC allergen technicians collected a four minute vacuum dust sample from the bedroom of each willing study participant that was 1 year old and older, vacuuming both the bed and the floor [5]. The dust sample was analyzed in a laboratory to test for 10 indoor allergens and endotoxins (cat antigen, dog antigen, mouse antigen, rat antigen, two forms of dust mite [*Dermatophagoides farinae, Dermatophagoides pteronyssinus*], two types of mold [*Alternaria alternada*, *Aspergillus fumigatus*], and two proteins in German cockroach [bla g 1, and bla g 2]). These home visits can be difficult, dangerous, and expensive; therefore we hope to identify appropriate survey questions capable of predicting home exposures in order to lessen the necessity of home visits.

There were three difficulties involved with the data: the size of the data set, missing entries in the data set, and the imbalance inherent in asthma data. The questionnaire was 930 questions (variables) and was completed by 10,348 persons (observations). Large data sets such as this are often too cumbersome for computational analysis; therefore, we reduced the number of questions considered by using combination of statistical methods and common knowledge pertaining to asthma and allergies. We first extracted the data corresponding to the 19 allergens and 10 exposures. The results from the tests for two of the exposures (*Aspergillus fumigatus* and German cockroach [bla g 2]) needed further analysis before they could be considered so the 10 exposure variables were lowered to 8. This decrease should not affect the model significantly because of the high cross-reactivity between the two mold species and the two cockroach species. We based our selection of the demographic, socioeconomic, dietary, and health- and home-related questions on factors known to trigger asthma and allergic reactions [1, 5, 6]. See Table B.1 for a complete list of the questionnaire variables used.

In addition to the data set being large, there were also many missing values. These omissions occurred occasionally because of laboratory measurement errors but most often were the result of the skip logic nature of the questionnaire. For example, if the questionnaire recipient answered that they did not own a pet, they would not have to answer the pet specific questions that directly followed. To ensure that the predictive models were based on as many observations as possible, we eliminated the variables in which more than a third of the observations had missing values.

The last difficulty that the data presented was the inherent imbalance due to the low percentage of asthma occurrence. Only 9% of the U.S. population has been diagnosed and is currently suffering from asthma. This small fraction can be difficult for statistical techniques to accurately predict and as a result, many statistical

algorithms would predict fewer cases of asthma than are actually represented in the data. To address this issue, we used Synthetic Minority Over-sampling TEchnique (SMOTE). SMOTE is an algorithm that adds new samples to the minority class, asthmatics, while also down-sampling the majority class, non-asthmatics, to provide a more balanced training data set [11]. The minority class is oversampled by creating "synthetic" examples. Such a synthetic example is obtained by selecting a random point along the line segment joining any two of the nearest neighbors in the minority class. The balanced training data was used to create the prediction models which were then tested using imbalanced test data.

3 The Approach

We first analyzed the NHANES 2005-2006 data to determine the association between allergen exposure, allergies, and asthma. This analysis was two-fold: initially a Structural Equation Model (SEM) was utilized to measure the correspondence between exposure, allergies, and asthma, then prediction models were generated to predict asthma based on the knowledge of allergies, exposures, and certain demographic information. After gaining an understanding of the relationship between exposure, allergies, and asthma, we developed prediction models for exposures in the home based on demographic, socioeconomic, dietary, and health- and home-related variables.

3.1 SEM Analysis

Initially, we used a Structural Equations Model (SEM) as an exploratory model on the raw data before we balanced it using SMOTE. The SEM assumed the 19 measured sensitivity levels, the measurements for the 8 exposures, and asthma diagnosis to be the *observed variables* which are affected by measurement error, while it defined three *latent variables* (i.e. not observed) which corresponded to asthma, exposure and allergies. Thus, the SEM can be written as a model with two layers. We have the *measurement layer* given as,

$$\mathbf{y}_n = \mathbf{v} + \Lambda \mathbf{z}_n + \boldsymbol{\varepsilon}_n, \quad \boldsymbol{\varepsilon}_n \sim \mathrm{N}(0, \Delta),$$

where $\mathbf{y}_n$, $\mathbf{z}_n$, $\mathbf{v}$, and $\boldsymbol{\varepsilon}_n$ are vectors, Λ is a matrix, and Δ is a diagonal covariance matrix. Specifically, $\mathbf{y}_n$ represents the observed variables for the nth subject, $\mathbf{z}_n$ contains the latent variables, $\mathbf{v}$ represents the population mean, and $\boldsymbol{\varepsilon}_n$ is the error. Additionally, we have the *structural layer*,

$$\mathbf{z}_n = \mathbf{a} + \Phi \mathbf{z}_n + \boldsymbol{\delta}_n, \quad \boldsymbol{\delta}_n \sim \mathrm{N}(0, \Sigma),$$

where $\mathbf{a}$ is the population mean, $\boldsymbol{\delta}_n$ is the random error for the nth subject, Φ is a matrix, and Σ is a covariance matrix. The SEM implies a set of covariances, and a maximum likelihood estimation (MLE) method is used to estimate parameters in the model. Additionally, we specified that exposure can influence allergies, and that asthma can be influenced by both allergies and exposures. Moreover, we assume all relationships to be linear and estimated the coefficients by fitting the model to the data.

Since asthma is very difficult to diagnose in children under six years of age, we first narrowed our focus to ages six and older. For the exposure variables, we used five exposures: the rat exposure was excluded because it is very rare, only the European dust mite was considered because of the extremely high cross-reactivity between the two dust mites, and the cockroach exposure was omitted in order to drastically improve the results. For the allergy variables, we chose six allergies: the five that corresponded to the chosen exposures and the cockroach allergy. A second SEM was developed for ages 6 to 18 since asthma tends to be most prevalent before adulthood. In order to create a more stable model for ages 6 to 18, we had to use different exposure and allergy variables. For both the allergies and exposures, we used five variables: Alternaria, cockroach, dog, European dust mite, and rat. The structural models that we obtained are included in section 4.1.

3.2 Prediction Models For Asthma

Two different statistical methods were used to model the occurrence of asthma, logistic regression using backward variable selection (in R and SAS) and random forests (in R). We fitted a logistic regression model to the data in order to predict the probability of asthma. For the model in R, we assumed a binomial distribution

for the response variable and utilized *logit* as the link function. Initially we used the 8 exposure variables, gender, ethnicity, age and income as well as the 7 allergy clusters found in Salo, et. al.[14] for the model. The income variable employed was the family poverty income ratio which represents the ratio of the family income to the appropriate poverty threshold based on geographic location.

Backward variable selection was employed to determine the variables used in the logistic regression models. We applied a 10-fold cross-validation technique which splits our training data into 10 equal subsets, fits the model to nine subsets, then verifies the model on the one that remains. We then use this model to estimate how accurately our model will perform. Details on both k-fold cross-validation and backward selection can be found in James et. al. [9]. Moreover, we used the fitted model generated via 10-fold cross-validation to predict on its corresponding test data, and we calculated the mean squared error (MSE) for accuracy.

Additionally, we addressed the question of how high the estimated asthma probability should be in order to conclude that the subject is asthmatic. A logistic regression prediction represents the probability of an individual having asthma, whereas the predictor (presence of asthma) is a yes/no variable. Thus a threshold had to be determined to decide whether or not a certain probability corresponds to an individual having asthma. Clearly, declaring a subject who has asthma to be nonasthmatic is a greater health risk than diagnosing a healthy subject with asthma. Consequently we wished to minimize the occurrence of false negative predictions (declaring an asthmatic to be nonasthmatic) more than the occurrence of false positive predictions (declaring a nonasthmatic to be asthmatic) although overall we wished to minimize both false predictions. We therefore developed a function based on a simple Monte Carlo simulation which determines an optimal cut-off point above which we consider a subject to have asthma. The function minimizes the expression $a + 2b$, where a denotes the false positive error frequency and b denotes the false negative error frequency. As a result, we calculated a cut-off probability of 31.5%, where if the prediction of asthma was below 31.5% the subject was non-asthmatic and if the prediction was above 31.5% the subject was asthmatic.

The logistic regression model in SAS was built using the 8 exposure variables, gender, ethnicity, age and income, but all 19 allergies were used instead of the 7 allergy clusters as in the logistic model in R. We chose the prediction cutoff value to be 0.437 by maximizing the Youden index; that is, asthma is diagnosed if the predicted probability is greater than 43.7%.

The second method we used to model the occurrence of asthma was classification and regression trees, utilizing the random forests algorithm. Random forests is an ensemble learning method that produces many classification trees rather than a single tree. One benefit to this method is its ability to handle large numbers of variables without deletion and to give a measure of variable importance which is useful for model selection. This type of algorithm is especially useful when backward variable selection is not appropriate. When we fit random forests with the training data set (approximately 2/3 of the data), we use bootstrapping to draw samples with replacement from the data set to grow the current tree. The remaining 1/3 of the data is used to get an unbiased estimate of classification error as trees are added to the forest. We used a 10-fold cross-validation to optimize the number of variables used at each split. For our random forest classification we created 500 trees with 70% of the data as training data and 30% of the data as test data. The random forest model was also constructed using the same variables that were used for the logistic regression in SAS. The results for all the asthma prediction models as well as a comparison of the asthma models can be found in section 4.2.

3.3 Prediction Models for Exposure

Another facet that we wished to explore was factors in the home environment that could predict home exposures. Specifically, we were hoping to discover a few common factors that seemed to contribute to exposures and allergens. To predict exposures in the home we first had to determine the questions from the survey related to the home that did not have a large number of missing entries. There were 71 variables that we used to predict the exposures; a complete list of the variables is included in table B.1. We then created three prediction models: Least Absolute Shrinkage and Selection Operator (LASSO) regression and random forest were performed using all 71 variables, and logistic regression with backward variable selection was done on a subset of this data containing 23 variables.

The first method we used was a classical logistic regression with backward selection using 23 variables from Table B.1; see table B.2 for a list of the 23 variables used. Restricting the number of variables was necessary because of the collinearity of several of the variables. Backward selection with collinear variables is

ineffective since it will often omit a variable that is important if it is linearly related to another. Therefore the 23 variables were selected by omitting the variables that were collinear. Because of time constraints, we used a 5-fold cross-validation to fit and test the models.

The second model that we developed employed a LASSO regression. LASSO regression is a linear regression with the lasso method for variable selection. That is, the lasso coefficients $\hat{\beta}_j$ $(j \in \{0, \ldots, p\})$ minimize the quantity

$$\sum_{i=0}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where the tuning parameter λ serves to control the so-called shrinkage penalty $\lambda \sum_{j=1}^{p} |\beta_j|$. A higher λ-value shrinks the parameters more, forcing more parameter estimates $\hat{\beta}_j$ to be equal to 0 and thus reducing the number of variables included in the model. We selected an optimal tuning parameter λ using 10-fold cross-validation. That is, we created a grid of possible values for λ and computed the cross-validation MSE for each value on the grid. Then we selected the λ that yielded the smallest cross-validation MSE. Finally, the model was refit using all available observations and the selected tuning parameter λ. We used the log-transformation of the exposure as the response variable $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$ in order to comply with the normality assumptions of linear regression. Whenever necessary, we applied the log-transformation on some of the predictor variables $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iN})^T$, also in order to ensure approximate normality.

The last technique we employed to predict exposure was a random forests classification. The random forests classification was performed using all 71 variables with 500 trees, 70% training data and 30% test data. A 10-fold cross-validation was again used as above to optimize the number of variables along each split. The results for the exposure prediction models can be found in the section 4.3 as well as a comparison of the prediction models.

4 Results

Below we have detailed the results of the prediction models for asthma and exposure. Additionally we have compared the results from each model to show which methods produced lower error rates and, as a result, better prediction rates. We will describe our first exploratory SEM analysis before moving on to the particulars of each of our prediction models.

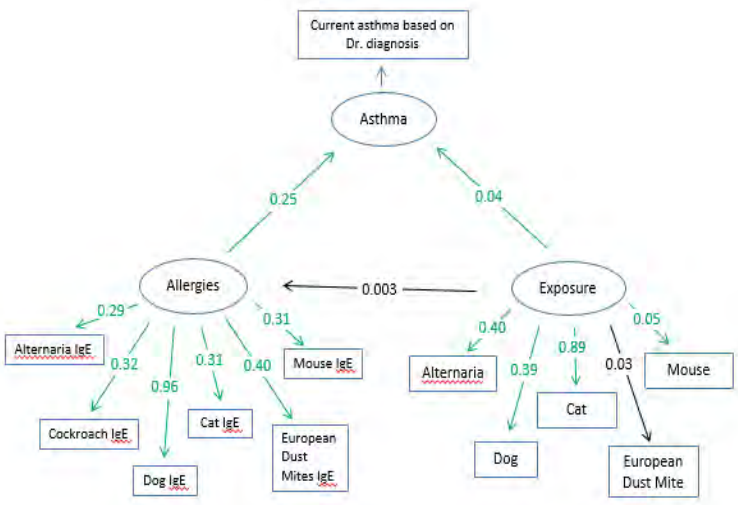


Figure 4.1: Structural Model for Ages 6 and Up

4.1 SEM Analysis

The SEM for ages 6 and older is available in Figure 4.1, where the coefficients and connecting lines are labeled in green if they are significant and black if they are not. We found that both allergies and exposures are significant in determining asthma; however exposure does not influence allergies significantly. Also, all of the six types of allergens are significantly important to measure allergies. Dog, european dust mites, cockroach are relatively most important. Cat, Dog, Alternaria are relatively important to measure the exposure.

The SEM for ages 6 to 17 can be seen in Figure 4.2. We found that allergies are significant in determining asthma, however the relationships between exposure and asthma, as well as allergies and exposure are not

Table 4.1: Confusion Matrices for Three Asthma Prediction Models

		Predicted asthma	
		negative	positive
Observed asthma	negative	1028	388
	positive	71	82

(a) Logistic Regression in R (accuracy: 70.75%)

		Predicted asthma	
		negative	positive
Observed asthma	negative	1054	330
	positive	71	64

(b) Logistic Regression in SAS (accuracy: 73.6%)

		Predicted asthma	
		negative	positive
Observed asthma	negative	1188	234
	positive	90	56

(c) Random Forests (accuracy: 79.34%)

statistically significant. All of the five types of allergen, especially Dog, are significantly important to measure allergies. Rat, Alternaria and European Dust Mites are relatively important when measuring exposure.

4.2 Prediction Models for Asthma

After examining preliminary results relating asthma, allergies, and exposure, we wished to investigate in more detail the relationships between asthma, allergies, and exposure. Consequently, we predicted asthma based on allergies and exposures as well as income, gender, age and ethnicity. The results for logistic regression in R and SAS, as well as the random forests classification are detailed below.

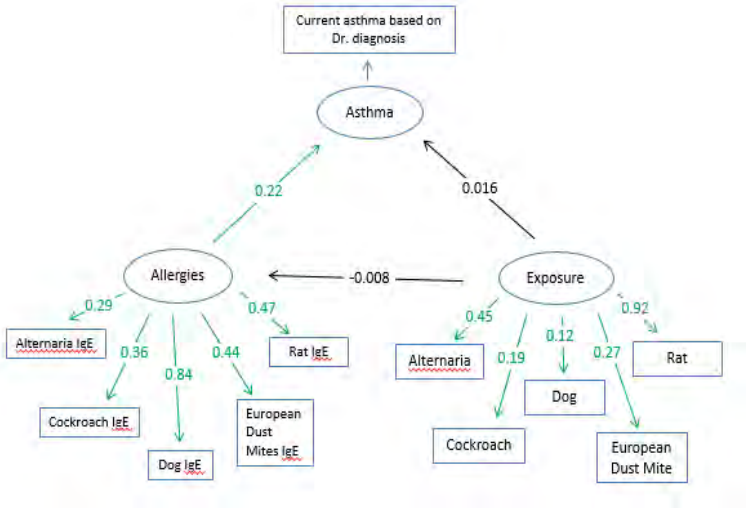


Figure 4.2: Structural Model for Ages 6-17

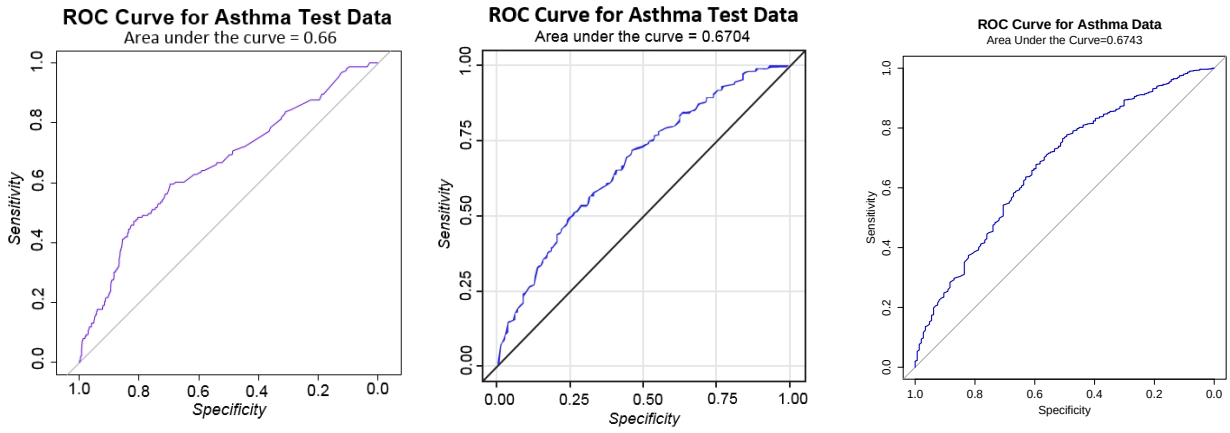
4.2.1 Logistic Regression using 7 allergy clusters in R

The logistic regression we performed used the 7 cluster variables for exposure rather than the 19 allergies. A table summarizing our findings for this model can be found in Table A.1. Additionally, we computed a confusion matrix which will compare the correct and incorrect predictions based on the observed data. The false positive prediction rate was only 27.4%, while the more concerning false negative rate was 46.4% This information is in Table 4.1a. ROC curves for the test data are given in Figure 4.3a where we calculated the area under the curve as 0.66. The accuracy associated with logistic regression was 70.75%.

What we found from our results is that exposure to dog and rat, and allergies such as food, mold, dust mite, roach, pet and rodent significantly increase the probability of asthma. Certain subgroups such as females, non-hispanic whites and non-hispanic blacks also had a greater probability of asthma. The model suggests a significant interaction between race/ethnicity and mold allergies. Figure 4.4 depicts this relationship. There we see that especially Hispanic/Latino ethnicities and Mexican-Americans have an increased probability of asthma in the presence of mold allergies, while non-Hispanic blacks and whites are less likely to have asthma. Interestingly, other races have a lower probability of asthma in the presence of a mold allergy than without the allergy.

Table 4.2: Error Rates for Three Asthma Prediction Models

	False Positive Rate	*False Negative Rate*	*Error Rate*
Logistic Regression in R	27.4%	46.4%	29.25%
Logistic Regression in SAS	23.84%	47.4%	26.4%
Random Forests	16.5%	61.6%	20.66%



(a) Logistic Regression in R (b) Logistic Regression in SAS (c) Random Forests

Figure 4.3: ROC Curves for Three Asthma Prediction Models

4.2.2 Logistic Regression using 19 allergies in SAS

We performed logistic regression in SAS using all 19 allergies. The individual level of detection for the different allergies was 0.25 kU/L. In R, we chose to transform these variables to 0 and 1, where a sensitivity less than 0.25 became a 0, and greater than 0.25 became a 1. For the SAS approach we decided to keep the numerical values as they were reported and not transform them to binary numbers. As with the logistic regression above, a results summary and confusion matrix can be found in Table A.2 and Table 4.1b respectively. The ROC curve is shown in Figure 4.3b with the area under the curve equaling 0.6704.

4.2.3 Random Forests

The confusion matrix for the random forests classification is shown in Table 4.1c with a false positive rate of 16.5%. Unfortunately our false negative prediction rate was 61.6%. The area under the ROC curve was 0.6743 which can be seen in Figure 4.3c. The accuracy of the random forest method was 79.34%.

We can also see which variables are the most important by looking at a Variable Importance Plot which is shown in Figure 4.5. According to the mean decrease Gini score, exposure to dogs and cats are the most important variables, followed by age, mouse exposure, and income.

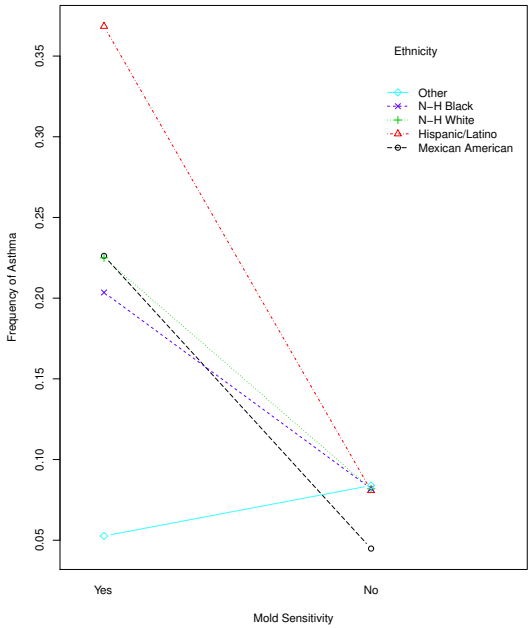
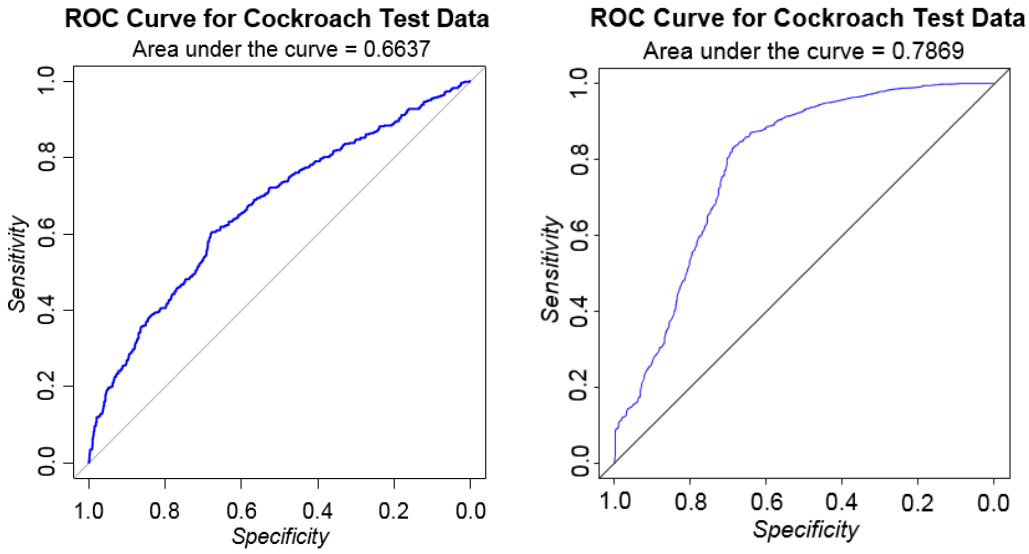


Figure 4.4: Interaction between mold allergies and ethnicity as predictors of asthma.

4.2.4 Comparison of the Methods



(a) Logistic Regression

(b) Random Forest

Figure 4.6: ROC Curves for Cockroach Exposure Prediction Models

Logistic regression and random forests are two very different methods with benefits and disadvantages to both. Logistic regression predicts the probability of asthma and classifies the predictions into groups of asthmatic or nonasthmatic based on a optimized cut-off, while random forest gives a definitive yes or no. Logistic regression is ineffective if there is a collinear relationship between some of the variables. On the other hand random forests can handle collinearity well and capture nonlinear relationships. The accuracy of random forests was superior at 79.34% compared to the logistic regression error rates, but the rate of false negatives was significantly higher for random forest.

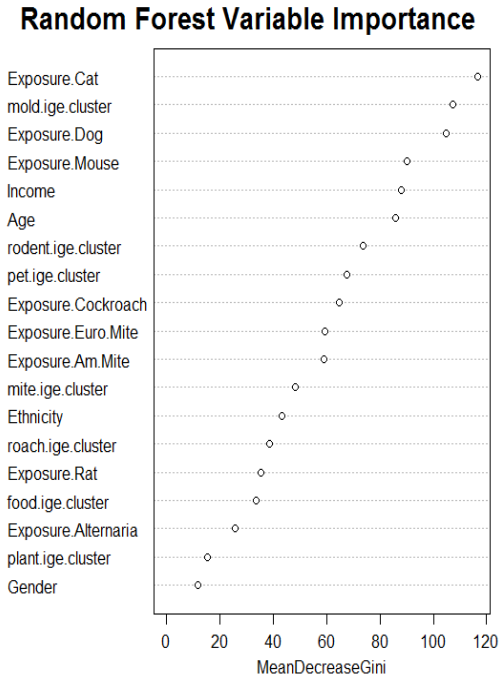


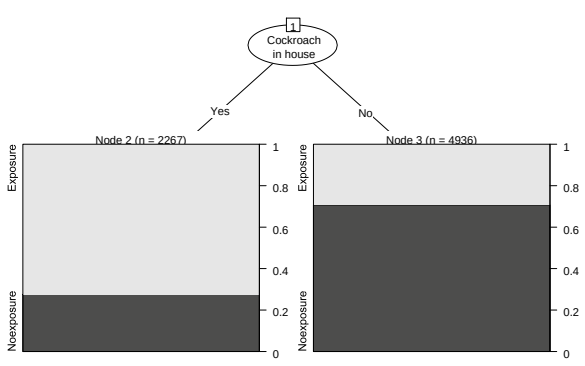
Figure 4.5: Variable Importance Plot for Random Forests Asthma Prediction

4.3 Prediction Models for Exposures

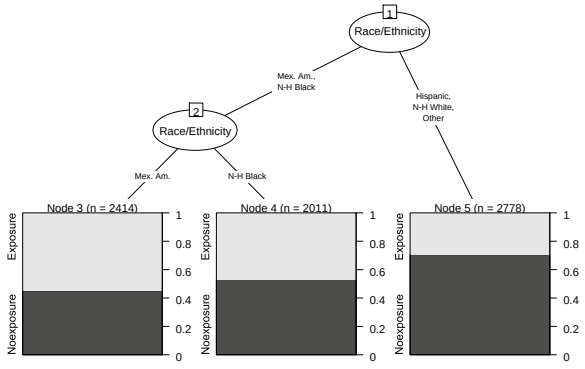
Another facet of interest was whether we could predict home exposures by identifying important factors in the home environment. While similar results could be computed easily in the future, we will only explain in detail the results for the cockroach exposure in this paper.

4.3.1 Logistic Regression

After performing a logistic regression with 23 variables using 5-fold cross-validation and backward selection, we obtained the important variables and their coefficients given in Table A.3. Having no impermeable pillow cover, or having 7 or more people in the household presented higher rates of exposure, while a detached one-family house, or a carpet rather than a smooth floor presented lower rates of exposure. The confusion matrix and error rates are available in Table 4.3. The overall accuracy is 66.83% with a false negative rate of 40.31%. The ROC curve is available in Figure 4.6a with the area under the curve equal to 0.6637.



(a) Tree for Cockroach Seen in Home



(b) Tree for Race/Ethnicity

Figure 4.7: Subgroup Trees for Exposure and Race/Ethnicity

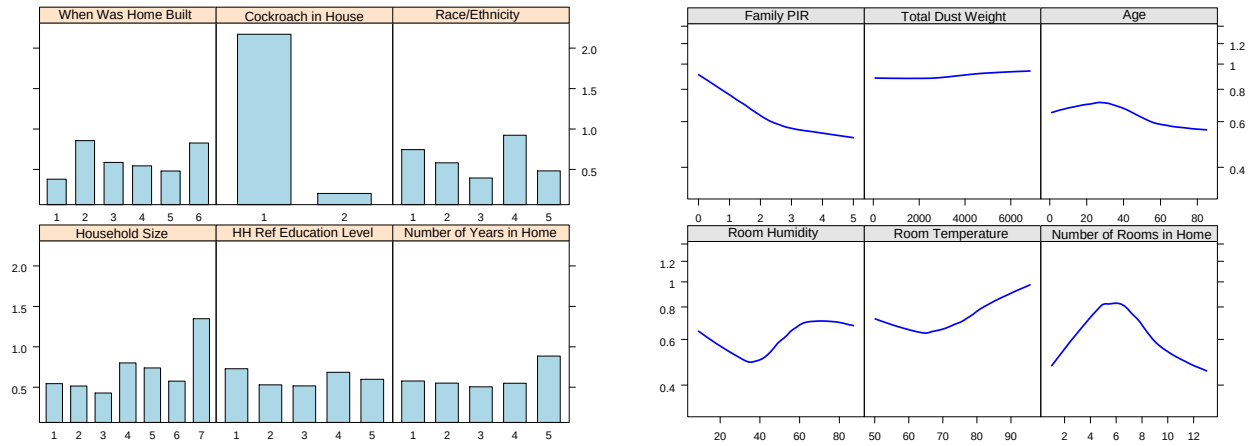
Table 4.3: Confusion Matrix Logistic Regression (accuracy: 66.82%)

		Predicted asthma	
		negative	positive
Observed asthma	negative	3570	1619
	positive	580	859
False Positive Rate 31.6%		*False Negative Rate* 40.31%	

4.3.2 Random Forests

We began our classification of cockroach exposure by examining a tree for exposure based only on whether or not a cockroach was seen in the home, Figure 4.7a, and based only on race/ethnicity, Figure 4.7b. Individuals who had seen a cockroach in their home had approximately 70% exposure rates as opposed to a 30% rate for those who had not seen a cockroach. Mexican-Americans had the highest rate of exposure at 55% followed by non-Hispanic black individuals at 45%. The lowest rates of exposures were among Hispanics (not including Mexican-Americans), non-Hispanic whites, and other races/ethnicities, whose rates were approximately 45%. Next we studied how multiple variables and their interactions influence cockroach exposure by building the classification tree seen in Figure 4.9. The most important factor was whether or not a cockroach has been seen in the home. Some other important factors were ethnicity, when the home was built, and how many people were in the home. Additionally, for Mexican-Americans and non-Hispanic black individuals, homes that were built before 1940 and homes in which they had lived for more than 10 years have a higher probability of cockroach exposure.

Our random forests model generated 500 trees similar to the one seen in Figure 4.9 and averaged the results of the trees. A subset of 12 important variables from the original 71 were selected based on the mean decrease Gini score. A bar chart containing 6 of these variables that have categorical data can be seen in Figure 4.8a, and partial dependence plots for the 6 continuous variables are contained in Figure 4.8b. The most important categorical variable was again seeing a cockroach in the home. Other important categorical variables were education, the size of the home, and ethnicity. Specifically, homes having 7 or more people, families living in the home for more than 10 years, and non-Hispanic black subgroups have a higher rate of exposure. For the important continuous variables, the most significant correlation is between income and cockroach exposure. Exposure levels decrease significantly with higher levels of income. The relationship between exposure levels and number of rooms in home is not linear: the exposure level peaks at 6 rooms. Moreover children have higher exposure level than adults and exposure levels are higher when the room humidity or temperature are at either extreme. An ROC curve can be seen in Figure 4.6b with the area under the curve equal to 0.7869.



(a) Categorical Variables

(b) Continuous

Figure 4.8: Important Variables Using Random Forests

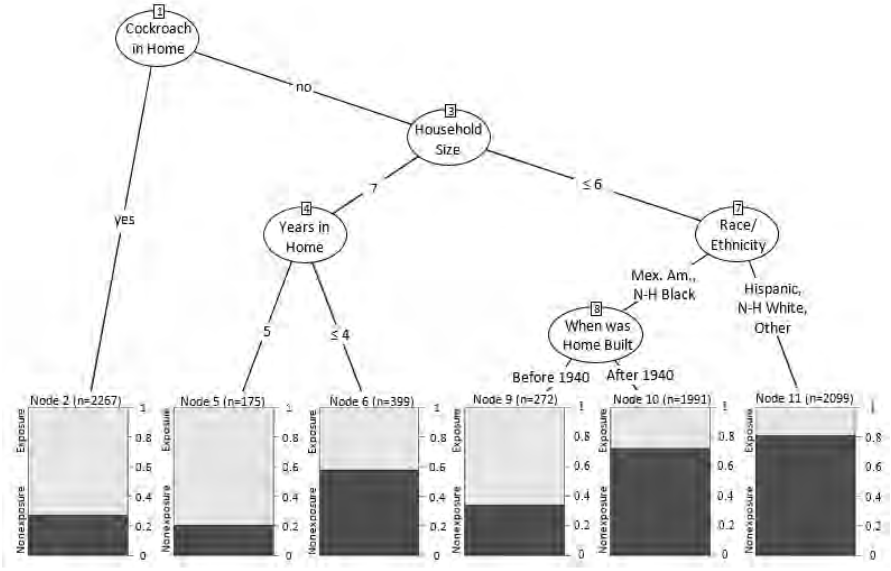


Figure 4.9: Tree for Cockroach Prediction

4.3.3 Lasso Regression

The results from the LASSO regression can be found in Table 4.4. Living in an apartment home as well as an increase in the amount of dust in the home increases the amount of cockroach exposure; while receiving a higher income and being a non-Hispanic white decreases the amount of cockroach exposure. Also, having seen cockroaches in the home indicates higher exposure levels.

4.3.4 Comparison of the Methods

While the LASSO regression model treated the amount of home exposure as a (continuous) numeric variable, the random forests approach and the logistic regression model converted the amount of home exposure into a bivariate categorical variable indicating low (insignificant) or high (significant) exposure. The models assumed exposure levels were low when the measurement was less than $1.77/\sqrt{2}$ U/g. Additionally, logistic regression predicted the probability of high exposure in the home, while the random forests model produced bivariate

Table 4.4: Lasso Results - Selected variables and the corresponding regression coefficients

Variable	**Type**	**Coefficient**
(Intercept)		-0.152580042
Income		-0.015684191
Ethnicity	Non-Hispanic White	-0.014172632
Type of Home	Apartment	0.012324286
Bed Surface Vacuumed	Blanket, Bedspread, or Comforter	0.009842256
Dust Weight		0.034204112
Have you seen cockroaches in your home?	No	-0.334296392
% deviance explained	MSE on test data	
0.157907	0.2000028	

estimates indicating whether home exposure was high or low. Therefore, the preferable method could depend on the exact objective of the researcher.

Nevertheless random forests classification has an advantage above the other two methods because, unlike the LASSO regression and logistic regression models, it captures nonlinear relationships between variables. In fact, Figure 4.8b clearly shows non-linear relationships between exposure and some of the predictor variables. This ability to recognize the nonlinearity between variables likely led to the random forests model having a smaller mean squared error (0.13) on the test data than the LASSO regression model (0.20). Additionally, the logistic regression model did not perform as well as other methods because there are linear relationships between some of the predictor variables. On the other hand, the random forest method and the LASSO regression are unharmed by collinearity.

In conclusion, we can say that the random forests method produced the best predictions for the analyzed data. Nevertheless, other factors should be taken into consideration when deciding on the method to use in a specific situation.

5 Summary and Future Work

It is generally accepted that asthma can be caused by allergies and environmental exposures and that environmental exposures also have an effect on the development of allergies, but the exact relationships are unknown. As described in the introduction, asthma puts a heavy burden on the quality of life of patients, as well as on the health care system. Therefore, it is important to identify the relationships between asthma, allergies and environmental exposures. Data analysis offers excellent techniques to help achieve this goal.

We considered three models for the prediction of asthma, two logistic regressions (where we used slightly different variables - allergies and allergy clusters) and a random forest model. A balanced training data set was created by oversampling the minority class (asthmatics) with the SMOTE algorithm. Model accuracy was tested on a separate test data set. The models showed many factors to be highly related to asthma and provided mathematical equations estimating whether a given subject is asthmatic or not. The three models agree that sensitivities to various allergens, home exposure to dog and to dust mite, female gender and race/ethnicity are important factors related to asthma. Additionally, both logistic regression models imply that there are significant interaction between ethnicity, allergies and asthma.

Since the mathematical models have shown that both exposure and allergy, are related to asthma, we considered a structural equation model (SEM) to explore how these three domains are interrelated. The SEM showed that pets, dust mite and cockroach allergies, as well as home exposure to pets and molds are the most important factors linking exposures and allergies to asthma.

The logistic regression, random forests and SEM models have shown that home exposures are important in predicting allergies and asthma. Unfortunately, the collection of household environmental data via home visits is difficult and expensive to obtain. Therefore, we considered predicting home exposures based on questionnaire data. We fitted three models, a logistic regression, a LASSO regression and a random forests classification for the prediction of home exposure to cockroach. The models provide mathematical equations to compute the amount of exposure based on questionnaire data. The three models show that factors such

as type of home, number of people in the household, size of the house, education level, income, ethnicity and room temperature have an influence on the amount of home exposure to cockroach.

Now that the mathematical models have identified which subgroups of the population are more prone to asthma (e.g. an allergic person or a female) and which subgroups are more prone to home exposure to cockroach (e.g. a family living in an apartment home or a low income family), we could move on to explore what extent such subgroups have on different patterns of exposure, allergy or asthma. A naive way toward that goal would be to consider interaction terms in regression models. Additionally, individual SEM or other models can be fitted to distinct subgroups including personal information and living conditions to further identify the association among these external factors, asthma, allergy and exposure.

A Logistic Regression Coefficient Tables

Table A.1: Logistic Regression for Predicting Asthma in R

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.7923	0.2163	-12.91	0.0000
Dog Exposure	0.3506	0.1573	2.23	0.0258
Rat Exposure	0.2611	0.1268	2.06	0.0394
Food Allergy Cluster	1.0566	0.1665	6.35	0.0000
Mold Allergy Cluster	2.5523	0.3527	7.24	0.0000
Dust Mite Allergy Cluster	0.6848	0.1269	5.40	0.0000
Cockroach Allergy Cluster	0.7423	0.1408	5.27	0.0000
Pet Allergy Cluster	1.0904	0.1298	8.40	0.0000
Rodent Allergy Cluster	2.1109	0.2634	8.01	0.0000
Female	0.2887	0.1146	2.52	0.0118
Hispanic (not including Mexican American)	0.5686	0.3561	1.60	0.1103
Non-Hispanic White	0.5897	0.1734	3.40	0.0007
Non-Hispanic Black	0.9932	0.1823	5.45	0.0000
Other Ethnicity	-0.0071	0.3466	-0.02	0.9837
Mold Allergy Cluster and Hispanic	-0.0357	0.9254	-0.04	0.9692
Mold Allergy Cluster and Non-Hispanic White	-1.1564	0.4261	-2.71	0.0066
Mold Allergy Cluster and Non-Hispanic Black	-1.9115	0.4095	-4.67	0.0000
Mold Allergy Cluster and Other Ethnicity	-1.5821	0.7658	-2.07	0.0388

Table A.2: Logistic Regression for Predicting Asthma in SAS

	Estimate	Std. Error	ProbChiSqr
(Intercept)	0.9322	0.2174	<0.0001
Dog Exposure	0.1117	0.0181	<0.0001
Cat Exposure	0.0955	0.0158	<0.0001
EU Dust Mite Sensitivity	0.2283	0.0343	<0.0001
Cat Sensitivity	0.3096	0.0507	<0.0001
Birch Sensitivity	0.1412	0.0451	0.0018
Aspergillus Sensitivity	0.7191	0.1600	<0.0001
Mouse Sensitivity	0.6979	0.1802	0.0001
Income	-0.1608	0.0566	0.0045
Gender	0.3048	0.0952	0.0014
Aspergillus and Ethnicity	-0.1192	0.0476	0.0122
Mouse and Ethnicity	-0.106	0.0467	0.0233

Table A.3: Logistic Regression for Predicting Cockroach Exposure

Variable	Type	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)		-0.3990	0.8725	-0.46	0.6475
Number of People in Household	2	0.2380	0.1909	1.25	0.2124
Number of People in Household	3	0.0495	0.1932	0.26	0.7976
Number of People in Household	4	0.0625	0.1940	0.32	0.7472
Number of People in Household	5	0.3302	0.1977	1.67	0.0949
Number of People in Household	6	0.3609	0.2220	1.63	0.1040
Number of People in Household	7 or more	0.6502	0.2135	3.05	0.0023
Ethnicity	Hispanic (Not Mex. Amer.)	0.1067	0.2441	0.44	0.6621
Race	Non-Hispanic White	-0.0563	0.1351	-0.42	0.6767
Race	Non-Hispanic Black	0.4210	0.1214	3.47	0.0005
Ethnicity	Other Ethnicities	-0.5883	0.2653	-2.22	0.0266
Type of Home	Detached One Family House	-0.7752	0.1563	-4.96	0.0000
Type of Home	Attached One Family House	-0.5262	0.1866	-2.82	0.0048
Type of Home	Apartment	-0.2626	0.1739	-1.51	0.1310
Type of Home	Dormitory	-1.0902	0.5731	-1.90	0.0571
Impermeable Pillow Cover	No	0.9355	0.6162	1.52	0.1290
Impermeable Pillow Cover	Pillow Not Present on Bed	1.2074	0.6481	1.86	0.0625
Floor Type	Medium/High Pile Carpet	-0.4105	0.1580	-2.60	0.0094
Floor Type	Smooth Surface	-0.0298	0.1289	-0.23	0.8173
FloorType	Carpet and Smooth Surface	-0.0592	0.2556	-0.23	0.8168
Education	9 - 11th Grade	-0.3044	0.1432	-2.13	0.0335
Education	High School Diploma	-0.6235	0.1434	-4.35	0.0000
Education	Some College or AA degree	-0.6346	0.1518	-4.18	0.0000
Education	College Degree or Higher	-0.4391	0.1861	-2.36	0.0183
Education	Refused to Answer	0.6226	1.4286	0.44	0.6630
Income		-0.1280	0.0384	-3.34	0.0009
When Home was Built	1978 - 1989	0.2006	0.1596	1.26	0.2088
When Home was Built	1960 - 1977	0.1212	0.1679	0.72	0.4705
When Home was Built	1950 - 1959	0.3317	0.1950	1.70	0.0890
When Home was Built	1940 - 1949	0.4322	0.2222	1.95	0.0517
When Home was Built	Before 1940	0.1652	0.1964	0.84	0.4003
When Home was Built	Refused to Answer	0.9463	0.9683	0.98	0.3284
When Home was Built	Don't Know	0.7335	0.1485	4.94	0.0000
Number of Years Lived in Home	1-2 Years	0.1513	0.1305	1.16	0.2461
Number of Years Lived in Home	3-5 Years	0.1836	0.1341	1.37	0.1711
Number of Years Lived in Home	6-10 Years	0.4878	0.1466	3.33	0.0009
Number of Years Lived in Home	More Than 10 Years	0.2943	0.1489	1.98	0.0481
Room Temperature (F)		-0.0193	0.0081	-2.37	0.0176

B Selected Demographic, Socioeconomic, Dietary, and Health- and Home-related Questions

Table B.1: Variables that Could Contribute to Home Exposure

Data Set Variable Name)	Variable Description
dmdhhsiz	Total Number of People in Household
dmdeduc.c	Education: High School or Less, Above High School
dmdhredu	Head of Household Reference Person Education Level
indfmpir	Family Poverty Income Ratio (PIR)
riagendr	Gender
ridageyr	Age
ridreth1	Race/Ethnicity
agqhay.c	Current Hay Fever
agqalg.c	Current Allergies
agq180.yn	Doctor Told Have Eczema
agq130.yn	Ever Had Itchy Rash in Least 6 Months
mcq010.yn	Ever Been Told You Have Asthma
mcqasth.c2	Current Asthma
rx.ar	Any Allergy Meds in the Last 30 Days
rx.ar.antihist	Any Antihistamine Allergy Meds in Past 30 Days
rx.ar.antiinfl	Any Anti-inflammatory Allergy Meds in Past 30 Days
rx.ar.antituss	Any Antitussive Allergy Meds in Past 30 Days
rx.ar.decon	Any Decongestant Allergy Meds in Past 30 Days
rx.ar.expect	Any Expectorant Allergy Meds in Past 30 Days
rx.ar.misc	Any Misc. Allergy Meds in Past 30 Days
rx.ar.par.steroid	Any Parenteral Steroid Allergy Meds in Past 30 Days
rx.ar.top.steroid	Any Topical Steroid Allergy Meds in Past 30 Days
rx.ar.vasocon	Any Vasoconstrictor Allergy Meds in Past 30 Days
rx.asth	Any Asthma Meds in Past 30 Days
rx.asth.antibody	Any Antibody Asthma Meds in Past 30 Days
rx.asth.antihist	Any Antihistamine Asthma Meds in Past 30 Days
rx.asth.antiinfl	Any Anti-inflammatory Asthma Meds in Past 30 Days
rx.asth.antituss	Any Antitussive Asthma Meds in Past 30 Days
rx.asth.β	Any β-agonists Asthma Meds in Past 30 Days
rx.asth.bronch	Any Bronchodilator Asthma Meds in Past 30 Days
rx.asth.decon	Any Decongestant Asthma Meds in Past 30 Days
rx.asth.expect	Any Expectorant Asthma Meds in Past 30 Days
rx.asth.leukoinh	Any Leukotriene Inhibitor Asthma Meds in Past 30 Days
rx.asth.misc	Any Misc. Asthma Meds in Past 30 Days
rx.asth.par.steroid	Any Parenteral Steroid Asthma Meds in Past 30 Days
rx.ecz	Any Eczema Meds in Past 30 Days
rx.exz.is	Any Immunosuppressant Eczema Meds in Past 30 Days
rx.ecz.top.steroid	Any Topical Steroid Eczema Meds in Past 30 Days
hoq011	Type of Home
hoq011.c	Type of Home: 1=Single, 2=Multifamily
hoq040	When Was Home Built (Categorical)
hod050	Number of Rooms in Home
hod060	How Many Years Family Lived in Home
hoq065	Home Owned, Bought, Rented, Other
aaxbdst	Bed Sample Status
aadbdsp	Bed Space Vacuumed (square inches)

Continued on next page

Table B.1 – continued from previous page

Data Set Variable Name	Variable Description
aadbdtim	Bed Vacuum Time (seconds)
aadbdtyp	Type of Bed
aaxbdsur	Bed Surface Vacuumed
aaxbdmat	Impermeable Mattress Cover
aaxfltyp	Type of Floor Covering
aaxbdplw	Impermeable Pillow Cover
aaxflst	Floor Sample Status
aadflsp	Floor Space Vacuumed (square inches)
aadfltim	Floor Vacuum Time (seconds)
aaxrmtmp	Room Temperature (F)
aaxrmhum	Room Humidity (%)
lbxdwt	Total Dust Weight (mg)
agq070	Removed Animal From Home Due to Allergy
agq090	Avoided Pets Because of Allergies
avoidpet	Removed and/or Avoided Pets in Home
hoq250	Do Animals Live or Spend Time in Home
hoq250.yn	Do Animals Live or Spend Time in Home, Currently
hoq260.dog	Do You Have an Indoor Dog, Currently
hoq260.cat	Do You Have an Indoor, Cat, Currently
hoq260.other	Do You Have an Indoor Small Furry Animal (not a dog or cat), Currently
hoq270.yn	Did Animals Live or Spend Time in Home, Past 12 Months
hoq070	Source of Tap Water
hoq080	Water Treatment Devices Used or Not
hoq230	Has Home Had a Mildew or Musty Smell
hoq240	Have You Seen Cockroaches in Your Home

Table B.2: Subset of Variables Used for Logistic Regression

Data Set Variable Name)	Variable Description
dmdhhsiz	Total Number of People in Household
dmdhredu	Head of Household Reference Person Education Level
indfmpir	Family Poverty Income Ratio (PIR)
riagendr	Gender
ridageyr	Age
ridreth1	Race/Ethnicity
agq180.yn	Doctor Told Have Eczema
mcqasth.c2	Current Asthma
rx.asth	Any Asthma Meds in Past 30 Days
hoq011	Type of Home
hoq040	When Was Home Built (Categorical)
hod060	How Many Years Family Lived in Home
aadbdtyp	Type of Bed
aaxbdmat	Impermeable Mattress Cover
aaxfltyp	Type of Floor Covering
aaxbdplw	Impermeable Pillow Cover
aaxrmtmp	Room Temperature (F)
aaxrmhum	Room Humidity (%)
agq070	Removed Animal From Home Due to Allergy

Continued on next page

Table B.2 – continued from previous page

Data Set Variable Name	Variable Description
avoidpet	Removed and/or Avoided Pets in Home
hoq250	Do Animals Live or Spend Time in Home
hoq270.yn	Did Animals Live or Spend Time in Home, Past 12 Months
hoq230	Has Home Had a Mildew or Musty Smell

References

[1] Guide to community preventive services. Asthma control: Home-based multi-trigger, multicomponent environmental interventions. `http://www.thecommunityguide.org/asthma/multicomponent.html`. Last updated: 05/01/2014.

[2] L. J. Akinbami, J. E. Moorman, C. Bailey, et al. Trends in asthma prevalence, health care use, and mortality in the United States, 2001-2010. NCHS data brief, National Center for Health Statistics (U.S.), May 2012.

[3] A. M. Branum and S. L. Lukacs. Food allergy among U.S. children: Trends in prevalence and hospitalization. NCHS data brief, National Center for Health Statistics (U.S.), October 2008.

[4] A. M. Branum and S. L. Lukacs. Food allergy among children in the United States. *Pediatrics*, 124(6):1549, 2009.

[5] CDC. *National Health and Nutrition Examination Survey (NHANES): Allergen Dust Collection Procedures Manual*. `http://www.cdc.gov/nchs/data/nhanes/nhanes_05_06/allergen_manual_06.pdf`.

[6] CDC. *You can control your asthma: A guide to understanding asthma and its triggers*. `http://www.cdc.gov/asthma/pdfs/asthma_brochure.pdf`.

[7] K. D. Jackson, L. D. Howie, and L. J. Akinbami. Trends in allergic conditions among children: United States, 1997-2011. NCHS data brief, National Center for Health Statistics (U.S.), May 2013.

[8] A. Jacquet. Innate immune responses in house dust mite allergy. *ISRN Allergy*, 2013.

[9] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*. Springer, 2013.

[10] C. C. Johnson, D. R. Ownby, E. M. Zoratti, et al. Environmental epidimiology of pediatric asthma and allergy. *Epidemiologic Reviews*, 24, 2002.

[11] M. Kuhn and K. Johnson. *Applied predictive modeling*. Springer, New York, 2013.

[12] J. E. Moorman, L. J. Akinbami, C. M. Bailey, et al. National surveillance of asthma: United States, 2001-2010, 2012.

[13] J. E. Moorman, R. A. Rudd, C. A. Johnson, et al. National surveillance for asthma – United States, 1980-2004. Morbidity and mortality weekly report, CDC, October 2007.

[14] P. M. Salo, S. J. Arbes, R. Jaramillo, et al. Prevalence of allergic sensitization in the United States: Results from the National Health and Nutrition Examination Survey (NHANES) 2005-2006. In press.

[15] K. B. Weiss and S. D. Sullivan. The health economics of asthma and rhinitis. I. Assessing the economic impact. *Journal of Allergy and Clinical Immunology*, 107(1):3–8, 2001.

Analysis of Self-Reported Health Outcomes Data from Web Based Media Sources

Fatena El-Masri[1], Karianne Bergen[2], Obeng Addai[3], Piaomu Liu[4], Shrabanti Chowdhury[5] Xin Huang[6]

Faculty Mentors: Mark Wolff[7], Kennith Lopiano[8]

Abstract

Historically, clinical outcomes have been derived from direct surveys of patients and formal reports of physicians. In the internet age, many patients are using the internet, including web-based media, social media and web forums, to share information and opinions about healthcare outcomes and adverse effects from drugs and medical devices. These spontaneous personal reports, describing drug-related or health-related data, provide both structured and unstructured information yielding a source of feedback about treatments and patient outcomes. New methods of analysis must be developed in order to make use of this growing body of data. One key element of this is assessing the usefulness or reliability of web-based information. Tools from text mining can be used to gather, validate and analyze the text data to determine its accuracy and reliability. We propose methods toderive relevant information from text gathered from web forums. More specifically, we consider methods for generating features including interactions among post authors, trending topics and sentiment analysis in web-forum posts, for analyzing and clustering posts related self-reported health outcomes.

1 Introduction

Due to the increasing number of unstructured text data on web-based media related to healthcare and patient outcomes, text mining techniques have become very important for analyzing data related to adverse effects of pharmaceuticals and medical devices for postmarking drug/device safety. Individuals are increasingly using the internet to find and share information related to healthcare. Twitter, Facebook, blogs, and web forums represent potential sources for unsolicited information that may be relevant to understanding possible unknown side effects, drug interactions or other adverse effects of drugs or medical devices. These data can be structured, but are often in the form of unstructured text. A key challenge in exploiting information from unsolicited web postings is the reliability and relevance of the data. Text posted on web forums may contain information that is inaccurate or biased. Therefore, before web-based information related to drugs and medical devices can be exploited to improve public health and patient safety, the challenge of assessing the accuracy and reliability of those data must be addressed.

Text mining is a set of tools designed to process text data and extract useful information from the text. SAS Enterprise Miner (EM) is useful software that can process raw text to create a table describing the statistics and features of the text. The text can be parsed and filtered and then topics can be derived from the corpus of documents. The features generated by EM can be used to classify or cluster the text to discover patterns in the data.

Here we propose new methods for mining text gathered from web-forums. We focus on methods that are not currently available in the EM software. We consider new features such as author interactions, trends analysis, sentiment analysis, and word homogeneity. These features may improve our ability to determine the relevance of posts for discovering adverse drug or device effects.

[1]School of Physics, Astronomy, and Computational Sciences, George Mason University
[2]Institute for Computational and Mathematical Engineering, Stanford University
[3]Department of Mathematics and Statistics, Youngstown State University
[4]Department of Statistics, University of South Carolina
[5]Department of Statistics, University of California Riverside
[6]Department of Mathematics, University of Texas at Dallas
[7]SAS Institute
[8]SAMSI

2 The Problem

Web-based media, including social media, forums, blogs, etc., contains a significant volume of spontaneous, self-reported health information. Some of these self-reported health data may provide useful information about patient treatments and outcomes, including previously unknown adverse effects of drugs or medical devices, to pharmaceutical companies, government regulators, doctors and patients. However, due to the large quantity of information being posted on the web-based media, algorithms are required to automatically filter the data collected from these media. For example, in the case of Avandia, a drug for diabetes, there were early indications based on web-forum posts that patients were concerned about increased risks of heart failure associated with the drug [1]. The ability to use information from these forums to detect these types of adverse events earlier than through traditional reporting systems, could have helped to prevent loss of life. In order to find useful signals, such as potential side effects of drugs, techniques for identifying relevant information among the many health-related postings. We seek to develop methods that can extract features from web-forum data that can be used to determine the usefulness of self-reported posts on health-related web-forums. The SAS Enterprise Miner (EM) software has tools that have been used to cluster health-related web posts based on text mining techniques [8]. We aim to extend the functionality of SAS EM to include new methods that provide additional information useful for analyzing data from web forums.

We approach this task by considering data about one particular medical device. We use a dataset that contains posts from four web forums in which users are posting about a medical device called the vagal nerve stimulator (VNS). Better understanding of the opinion and outcomes related to the VNS device will allow for improved regulatory monitoring and better medical decision-making. Data for 28279 forum posts by authors with 4173 unique usernames are included in this dataset. The data includes the Author, Title, Timestamp, Text, Web Source, and a few other fields, for each forum post (or "document").

3 The Approach

In order to determine if a given post contains useful information related to adverse effects from a drug or medical device, we aggregate a set of features associated with the document. Features are derived based on the text in the post, temporal patterns, and the author of the post.

From the web-forum data, we can easily determine the length of each post, in terms of number of words. We can associate each post with its author and also identify the frequency and total number of posts written by that author (unique username).

Using SAS EM software, we can determine the topics mentioned in a particular post. Previous work with EM has also used the homogeneity of words the documents to identify single authors posting under multiple usernames, and to segment authors into "buckets" with certain characteristics, such as "salesmen," "moderators," or "narcissists".

We also propose methods to generate new features from the documents that are not currently available in EM. We describe a burst detection method to identify "trending topics" within the corpus of forum posts. We apply methods for sentiment analysis to determine the positive or negative sentiment of the post. We investigated the utility of a method for determining the degree of interactions among different authors in the forums.

Our approach to determine the usefulness of health-related forum posts is to collect all of these features into a table for each document. This table can then be analyzed using classification or clustering methods to determine which documents contain relevant information.

3.1 SAS Text Miner

SAS Text Miner a powerful tool for extracting valuable information from a set of unstructured documents. Using the SAS Text Miner software, one can explore the relationships within a set of documents and make valuable conclusions about those documents.

Here we review the existing features in SAS Text Miner and identify additional methods useful for text mining that are not currently implemented in SAS TM.

3.1.1 SAS Text Miner

The following features are currently included in SAS TM.

- **Import File:** SAS Text Miner can import a variety of file types including `csv`, `jmp` and `xlsx`. The first step in the process is to specify the type of the file and the directory of the file.
- **Text Parsing:** In this stage, the sentences in in a document are separated into individual words, which are labeled according to their part of speech (e.g. nouns, adverb, adjective, abbreviation) according to parameters set in the Text Parsing Node. There are other options such as the Language(English or Spanish) used in the data.
- **Text Filter:** The Text Filter follows the Text Parsing node, and makes a table of terms in the document, the frequency and weight of each term, the number of documents in which a term appears, an option to keep a term for further analysis.
- **Text Topic:** The Text Topic node is very powerful in the sense that it identifies and group terms together as a topic using a dimensionality reduction method. A number of documents are identified by the topic and grouped together under that topic. Each document in the corpus is represented in terms of a small number of topics contained in the document. The dimension reduction tool used here is the Singular Value Decomposition(SVD) and details of the the process are stated in the paper by Russ Albright [10].

As pointed out earlier, SAS Text Miner uses the SVD to cluster documents into a specified number of topics. Recent research have shown that, negative values in some of the dimensions generated by the SVD are less meaningful in terms of capturing the base latent semantics of the document corpus. An alternative clustering and dimensionality reduction method for nonnegative datasets, the Nonnegative Matrix Factorization, is introduced in section 3.2.

3.1.2 Document Corpus Partitioning Methods

In this section, we describe some of the techniques which can be applied in SAS to partition or cluster the document corpus into groups of "useful" and "not useful" documents.

- **Document Size:** The number of terms in a document can provide information about the usefulness of that document. A patient who wants to share an experience with a product on the market may need more than 5 words to describe his/her situation. In other words, depending on the product or problem in question, any "useful" post may need more than some specific number of words to describe an experience. Using this approach may help eliminate posts which are too short to contribute any valuable information.
- **Document Homogeneity:** We describe a procedure which can eliminate posts written by bots or angry authors who only wants to be heard by posting similar documents. Using the author frequency results from SAS Text Miner, we can sample (about 5 percent) of the documents from an author a large number of posts. Next, we use the terms which make up this 5 percent to check the similarity in the rest of the corpus. Using this method, we are able to detect bots and authors who write the same posts, including posts by a single author under multiple usernames.

  There are cases in which an author may have a high frequency of posts but if he is addressing questions of other authors, this procedure will not eliminate the documents of such an author.

3.2 Nonnegative Matrix Factorization

An alternate technique for clustering data is called nonnegative matrix factorization (NMF). Nonnegative matrix factorization decomposes nonnegative matrix X into the product of two reduced-dimension nonnegative matrices.

Consider an $m \times n$ matrix X for which every entry is nonnegative, where the rows represent m features and the columns represent n data samples: $X_{ij} \geq 0$, $\forall i, j$, and $X \neq 0$. Then the NMF for X takes the form

$$X \approx WH, \quad \text{where } W \in \mathbb{R}_+^{m \times k} \text{ and } H \in \mathbb{R}_+^{k \times n}, \text{ and } k < \min(m, n).$$

W is a dictionary of positive basis elements or "metafeatures" and H is a matix of coefficients. The matrix W can be viewed as lower dimensional basis for X, and thus NMF can be used a dimensionality reduction method.

W and H can be computed by solving the non-convex optimization problem

$$\begin{aligned} \underset{W,H}{\text{minimize}} \quad & \|X - WH\|_F^2 \\ \text{subject to} \quad & W \geq 0,\ H \geq 0 \end{aligned}$$

for which specialized solvers have been developed[7].

The advantage of NMF lies in the nonnegativity of the factor matrices W and H. Let w_i be the ith column, or "metafeature", of W. The for any data sample x_j, the jth column of our original data set X, we can approximate x_j as a linear combination of the "metafeatures" with the coefficients found in the jth column of matrix H:

$$x_j \approx \sum_{i=1}^{k} w_i h_{ij}$$

Both w_i and h_{ij} are positive for all i and j, so data sample x_j is a positive linear combination of nonnegative basis vectors in W. This makes NMF more interpretable than other matrix decompositions since each data sample can be build by adding up elements from the basis W of positive elements. The singular value decomposition (SVD) can be difficult to interpret due to cancellation of positive and negative values appearing in the singular vectors.

The dictionary matrix W from NMF can be used to partition data into k clusters. We use a training dataset X to compute the dictionary W. Then for any data sample x, we can compute the set of nonnegative coefficients $h \in \mathbb{R}_+^k$ of x in basis W, that is $x = Wh$, and associate x with the cluster $c = \arg\max_\ell(h_\ell)$ corresponding to the largest coefficient in h. The number of clusters k can be selected using the model method of Brunet et. al. [3]. This method uses consensus clustering to select a value of k for which the partition of the data remains relatively consistent for different initializations of the NMF solver.

3.3 Sentiment Analysis

Sentiment Analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information from source materials [9]. Generally speaking, Sentiment Analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.

Our goal is to distinguish between "useful" forum posts and "useless" forum posts. Among useless forum posts, some contain too many emotional words. The sentiment of a post may be a helpful feature for determining its usefulness because emotional posts may be associated with author bias. For example, posts such as

- *"omg, that's too funny, lol!"*, or
- *"I am happy to add Cheryl Jones to our list of ETP Champions. Please help support her efforts. Let her know you appreciate this!"*, or
- *"Hey Gurl That's messed up of your parents to be the way they're acting with you.... He's never offered me money for my Meds or anything and I pay out of my pocket too cause I no longer have insurance but just worry about yourself since they're only worrying about themselves !!!"*

contain many "sentiment words" and their content does not contribute useful information with respect to adverse effects of the VNS device. Since there will be posts in which authors use emotional words but they are

talking about some relevant information, we treat the sentiment of the post as just one feature to distinguish "useful" posts and "useless" posts.

To determine the sentiment score for an individual post, we count the number of positive words in the post, NP, and the number of negative words appearing in the post, NN. Then we can define the sentiment score, S, as

$$S = NP - NN$$

The sentiment score will be positive if the post contains more positive words than negative words and negative if the post contains more negative words. The magnitude of the score will tend to be larger for documents that contain more "emotional" language.

3.4 PageRank for Author Interaction

Toward the goal of filtering out authors who post unreliable, irrelevant, or unhelpful information on web forums, we propose calculating a score for each author based on his interaction with other users on web forums. This approach will allow us to detect and filter out posts written by authors who are posting frequently but receive few replies or reposts, suggesting that other members of the community do not find their posts relevant. As a starting point, we can measure the quality of each author's interaction with other forum users based on the structure of the network replies or reposts among authors.

Consider a corpus of documents with posts written by M authors. Then we can create the $M \times M$ author adjacency matrix A such that

$$A_{ij} = \begin{cases} 1 & \text{if author } i \text{ replies to a post written by author } j \\ 0 & \text{otherwise} \end{cases}$$

The matrix A represents a directed graph in which each vertex corresponds to an author in the corpus. A non-zero value in the matrix A in entry (i, j) represent an edge from vertex i to vertex j, indicating that author i responded to a post written by author j.

We can use this graph to determine clusters of authors or the most important authors within the corpus of posts. To determine the most important authors in the post we use the *PageRank* algorithm[2], a method for link analysis which computes the relative importance of each vertex (author) in the graph.

The PageRank algorithm was originally used to rank the importance of webpages based on their link structure. The algorithm is based on a "random surfer" model of the web. An internet-user at webpage i will browse the web by clicking on a link to a new page j chosen at random according to probability distribution p_i based on the links on page i. With probability α the user will jump to a webpage selected uniformly at random among all possible webpages. Similarly, if the user reaches a page with no outgoing links, he will jump to a page chosen at random from among all webpages. The PageRank algorithm defines the relative importance of webpage i to be the probability that the user lands on webpage i under this model. Mathematically, we can represent the user's browsing as a random walk on a Markov chain with transition matrix P, and compute the steady-state distribution $x = xP$ to obtain the relative importance x_i of webpage i.

In our application, we are interested in the interaction among authors rather than the link structure of webpages. We wish to rank the relevance of each author's posts by considering the linkages among forum posts based on replies or reposts. Using the author adjacency matrix A, we can compute the transition matrix P for our author ranking problem.

Consider the author adjacency matrix A defined above and let $R_i = \sum_{j=1}^{n} A_{ij}$ be the number of replies posted by author i, then the transition matrix P is given by

$$P_{ij} = \begin{cases} \frac{1}{M} & \text{if } R_i = 0, \quad \text{i.e. if author } i \text{ has not posted any replies} \\ (1-\alpha)\frac{A_{ij}}{R_i} + \frac{\alpha}{M} & \text{if } A_{ij} = 1, \quad \text{i.e. if author } i \text{ replied to a post by author } j \\ \frac{\alpha}{M} & \text{if } A_{ij} = 0 \text{ and } R_i > 0 \quad \text{i.e. if author } i \text{ posted a reply, but not to author } j \end{cases}$$

Using this transition matrix, we can rank each author's level of interaction in the forum community by solving for the steady-state distribution, x, such that

$$x = xP.$$

We can compute the steady-state distribution x using the power iteration method. Letting $x_i^{(0)} = \frac{1}{m}\ \forall i$, we iterate

$$x^{(k+1)} = x^{(k)}P \qquad \text{until convergence of } x^{(k)}.$$

The interaction score for author i is equal to x_i. Since authors who receive more replies or reposts have higher interaction scores, we believe that these authors are likely to be posting information that are more relevant or useful to other members of the forum community. In contrast, authors who received relatively few replies will have lower interaction scores.

3.5 Burst Detection for Topic Trending

We study the temporal aspect of a topic on a forum like Epilepsy. We refer to a two-stage weighted automation model that can be extended to an infinite state model [11].

In text mining analysis one of the important tasks is to identify patterns or structures in the stream of messages that can improve the understanding of large volume of messages. Documents can be easily classified into different categories by topic using basic clustering methods. Usually our data contains the arrival times of the documents. The topics contained in the messages appear, rise high in intensity and then dies down again over time. When a particular topic appears, corresponding features that define that topic tend to appear more as well – indicating a burst of activity. Even the documents related to a single topic arrive at different rates over time and thus the time intervals can be split into more localized periods which correspond to the bursts of different intensities depending on the rate of the message arrivals. Here our goal is to present a method to model such "bursts" so that they can be detected efficiently and hence the underlying content of the documents containing trending topics can be analyzed further. The basic concept of this method is that the bursts correspond roughly to the time points at which documents arrive more frequently reducing the interval gap between their arrivals, probably once in a week to once in an hour or minute. The rate of arrivals show frequent alternations of rising and fading away and thus this method is based on analyzing the inter arrival gaps of the consecutive documents to identify the large number of short bursts which are contained in a long burst.

Here the model is based on a discrete-space Markov process which at any point of time can be in one of its possible set of states which correspond to the different intensity levels and the emission of the messages at different rates depend on the state where the process is. Burst occurs whenever there is a transition from a lower state to higher state. The frequency of such transitions can be controlled by considering a cost factor with every transition from lower to a higher state and thus very short bursts of tiny intervals can be prevented and identification of long persisting bursts is easier. The bursts associated with state transitions form a hierarchical structure, where a long burst of low intensity contains inside it several small bursts of higher intensities. Considering the simplest randomized model, the inter arrival times are generated from the exponential distributions: $f(x) = \alpha e^{-\alpha x}, x > 0, \alpha > 0$ is the rate of arrival of messages.

The basic model considers only two states of the Markov process q_0(low intensity) and q_1 (high intensity) and the transition probability from one state to the other is independent of previous arrivals or state changes. This gives the transition probability matrix as

$$P_{0,1} = \begin{pmatrix} p_{0,0} & p_{0,1} \\ p_{1,0} & p_{1,1} \end{pmatrix}$$

, where $p_{i,j} \in [0,1], i = 0,1; j = 0,1$.When the process is in state q_0 , documents arrive according to the exponential distribution with parameter α_0 and with α_1 when in state q_1 with $\alpha_1 > \alpha_0$. Suppose there are $(n+1)$ documents and their arrival times which give n inter-arrival gaps $x = (x_1, x_2, \cdots, x_n)$, and we assume $x_i > 0, \forall i$. The density function of the state sequence $\mathbf{q} = (q_{i1}, q_{i2}, \cdots, q_{in}), i \in \{0,1\}$ is given as

$$\begin{aligned} f_q(x_1, x_2, \cdots, x_n) &= \prod_{t=1}^{t=n} f(x_{i_t}), \\ \text{where } f(x_{i_t}) &= \alpha_0 \exp(-\alpha_0 x_{i_t}), \quad i = 0 \\ f(x_{i_t}) &= \alpha_1 \exp(-\alpha_1 x_{i_t}), i = 1. \end{aligned}$$

Now if b denotes the number of state transitions in the sequence, then the prior probability of the state sequence $\mathbf{q}$ is given as: $p^b(1-p)^{n-b}$, where p is the probability of transition. Assuming the process starts in the state q_0 , the posterior probability of $\mathbf{q}$ conditional on x is given as:

$$P(q|x) \propto p^b(1-p)^{n-b} \prod_{t=1}^{t=n} f(x_{i_t}).$$

Finding a state sequence $\mathbf{q}$ maximizing this posterior probability is equivalent to finding one minimizing the cost function:

$$-\ln P(q|x) \propto -b\ln(p) - (n-b)\ln(1-p) + \sum_{t=1}^{n} -\ln\left(f_{i_t}(x_{i_t})\right)$$

This is equivalent to minimizing the cost function:

$$c(q|x) = b\ln\left(\frac{1-p}{p}\right) + \sum_{t=1}^{n} -\ln f(x_{i_t})$$

, since other terms are independent of the state sequence. Here the two terms of the cost function balance out well to track the global structure of bursts in the gap sequence. This method can be easily extended to more than two states model (in fact to an infinite states model). The inter arrival gaps are assumed to be generated from $f(x) = \alpha_i e^{-\alpha_i x}, x > 0$, where $\alpha_i = \hat{g}^{-1}s^i$, and $\hat{g}$ is the gap size if the messages were spaced evenly on the time line. Let $\tau(i,j) = (j-i)\gamma\ln(n)$ be the cost associated with a transition from q_i to $q_j, j > i$. Then, the optimum state sequence is obtained by minimizing the cost function

$$c(q|x) = \sum_{t=0}^{n-1} \tau(i_t, i_{t+1}) + \sum_{t=1}^{n} -\ln f(x_{i_t})$$

Here s is the scaling parameter on which depends the tracking of the time intervals of bursts and the process changes state depending on the parameter γ since it is associated with the cost factor.

The optimal state sequence obtained from optimizing the cost function will give us the hierarchical structure of the set of bursts. The burst of intensity j will be an optimal interval where the state sequence is in state j. More formally, it is the interval $[t_1, t_2]$ such that $i_{t_1-1} < j, i_{t_2+1} < j$, and $i_{t_1}, \cdots, i_{t_2} \geq j$. So the bursts show a nested structure where the burst of higher intensities are contained in those of lower intensities, i.e., burst of intensity j contains one or more sub intervals that are bursts of intensity $j+1$ each of which may further contain sub intervals that are bursts of intensity $j+2$ etc. Hence the optimal sequence can be transformed to a nested structure of bursts which can be captured by a rooted tree diagram where the root of the tree corresponds to a single burst of lowest intensity over the whole interval [0, n]. Once we get the bursts of different intensities for each topic we can use them to identify if the documents contain any trending topic or not. Long bursts of higher intensities give possible indications of trending topic in the documents.

3.6 Decision Tree classification

After we have aggregated a set of features for each document, we can cluster the data using an unsupervised learning method such as NMF. If annotated data are available, we can obtain better results for predicting which posts contain useful content and which are irrelevant by applying a supervised machine learning method. First, we must obtain examples of posts in each class, "useful" and "useless", annotated by a domain expert. A supervised machine learning method should be selected based on the properties of the data set. The features we have proposed include both categorical variables, such at the document topics or author, and numerical variables such as the length of the post or sentiment score. We also observe that for text data collected from internet forums, the data is often incomplete or missing, resulting in missing features for some documents. Based on these characterstics of our data and feature set, we believe that decision-tree learning algorithms are most suitable for this application, as they can easily handle both categorical and numeric variables and can handle missing data using surrogate variables. Previous work that applied a decision tree classifier using a subset of our proposed features performed well, so we expect that extending the feature set to include the information about sentiment, trending topics, and author interactions will improve performance.

4 Computational Experiments

4.1 Sentiment analysis on VNS forum data

The goal of sentiment analysis is to score each post based on a sentiment function. Generally speaking, we want to use lexicons which contain positive and negative words separately. We match the words in each post against the words in the positive word lexicon and in the negative word lexicon, to obtain the number of positive words and number of negative words occurring in each post. Then we score each post based on our defined sentiment score function.

Since raw (unprocessed) posts consist of sentences, we need to clean the data before we can apply sentiment analysis. Cleaning the posts involves removing punctuation, characters and digit, and then splitting each sentence into a list of its component words.

Once we have obtained the list of component words for each post, we match the words in a post to the words in the positive and the negative lexicons. This allows us to compute the sentiment score:

$$S = NP - NN$$

where NP is number of positive words appearing in each post, and NN is number of negative words appearing in each post.

Based on this procedure, we calculated the scores for each post in the corpus. Examples of the text and scores for three of the posts are given in the table below. We can see that first two authors express words associated with positive emotions so their scores are positive, while the third author uses words associated with negative emotions so his score is negative.

Post	Score	Text
Number 2	2	*omg, that's too **funny**, **lol**!*
Number 890	3	*I am **happy** to add Cheryl Jones to our list of ETP Champions. Please help **support** her efforts. Let her know you **appreciate** this!*
Number 933	-4	*Hey Gurl That's messed up of your parents to be the way they're acting with you.... He's **never** offered me money for my Meds or anything and I pay out of my pocket too cause I no longer have insurance but just **worry** about yourself since they're only **worrying** about themselves !!!*

4.2 PageRank for Author Interaction

The author interaction scoring method was tested on two data sets. The first data set is the VNS data provided by SAS Institute and the second data set was an artificially generated data set modeling different classes of authors. Previous analysis of the VNS data set indicated that many of the authors who post very frequently are not posting relevant information. However, using a fixed threshold to remove posts by frequent authors may eliminate some authors who post helpful content and are highly active in the forum community. Therefore, for both data sets we seek to test whether the author interaction score can aid in distinguishing between active authors who post relevant content and those who are spamming forums with irrelevant posts.

4.2.1 VNS forum data

For this experiment, we used the subset of forum posts in the VNS data set for which both the post "Author" and "Title" were available. Out of the 28279 forum posts and 4173 post authors contained in the VNS data set, the Author/Title data was available for 4480 forum posts written by 2111 authors.

For the VNS data we do not have precise information about which posts are reponses to other posts. In order to test PageRank for author interaction scoring, we used the titles of the posts to group all posts with the same titles into threads of related posts. For example, three posts titled "Re: Deja vu feeling" are assumed to belong to a single thread (discussion) within the forum. The 4173 posts were grouped into 1861 threads, which range in length from 1 to 50 posts. To determine author interactions, we assume that all author interactions within each thread are symmetric; if author i and author j post in a common thread, we record this interaction as both "author i replied to author j" and "author j replied to author i". This results in 32972 total links between authors, which were used to construct the adjacency matrix A_{vns} and the transition

matrix P_{vns} with PageRank parameter $\alpha = 0.02$. The relative author interaction scores, x_{vns} was computed using the power iteration, which converged in fewer than 10 iterations. Finally, the author interaction scores were rescaled so the lowest and highest scores were 0 and 1, respectively.

The rescaled author interaction score for each author is plotted against the number of posts in Figure 1. Based on these two variables, we can assign the authors to three groups; there are authors with high interaction and many posts (shown in green in Figure 1), authors with low interaction and many posts (red), and authors with low interaction or few posts (blue). This suggests that rather than using a single threshold to eliminate posts by frequent authors, we may be able to get a better segmentation of documents by considering the number of posts by the author in combination with information about interactions with other posters in the forums. For example, if we eliminated authors in this data set subset who posted more than 15 times, we would be able to eliminate all of the low-interaction, frequently posting authors (red), but we would also eliminate all of the high-interaction, frequently posting authors (green) who may be posting relevant content. If we were to raise that threshold to a higher value, for instance 80, we would eliminate the two authors with the most posts, but we are leaving many of the less interactive authors with above average number of posts. However, if we only eliminate post by authors who both have low interaction and many posts we will remove many irrelevant posts while retaining the useful posts of high-interaction, frequently posting authors.

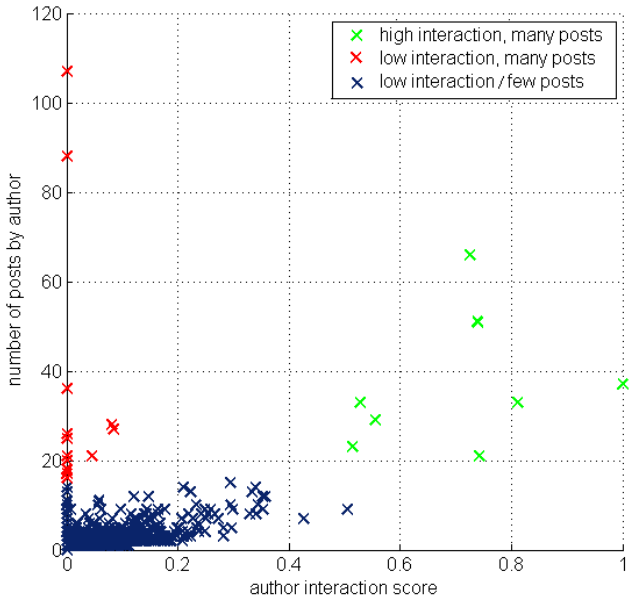


Figure 1: Plot of the author interaction score, rescaled to the interval $[0, 1]$, against the total number of posts by the author for subset of VNS forum data. Each 'x' in the plot represents one author and is color-coded based on the author interaction and post frequency.

4.2.2 Simulated forum data

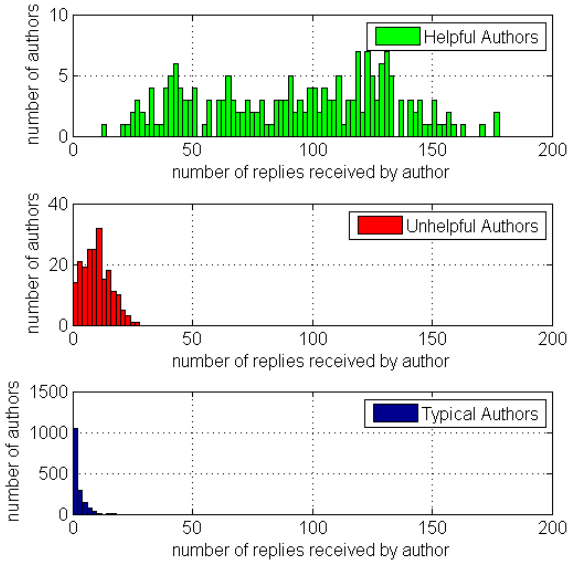
The experiments on the VNS data indicate that PageRank for author interaction scoring may be useful to distinguish between different types of authors who post relatively frequently: those who contribute relevant posts that receive many replies and those who write irrelevant posts that receive few responses. However, those results are based on only a subset of VNS forum data with incomplete information about the author interactions. Furthermore, we do not have information to confirm that there are in fact two distinct classes within the group of frequent posters. Therefore, to demonstrate the performance of the method, we used a simulation to generate an artificial data set that includes complete information about author characteristics and linkage between individual posts.

The simulated data includes 2000 authors, 10 percent "helpful" authors, 10 percent "unhelpful" authors, and 80 percent "typical authors". The posting patterns of each group varies, in terms of number of posts and number of forums in which posts appear. The groups also vary in terms of how likely it is that others will

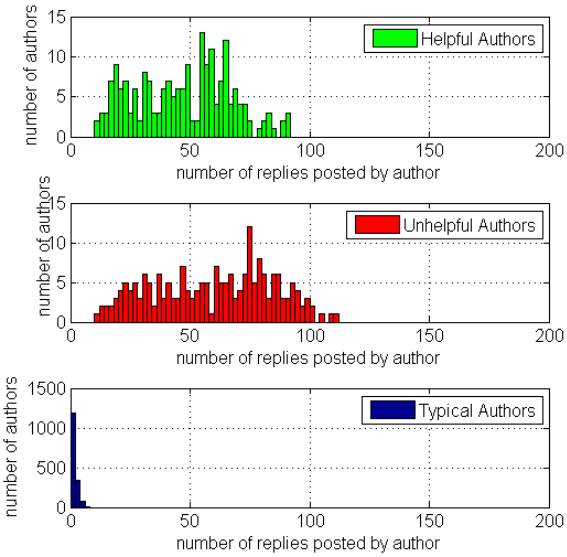
reply to their posts, with "helpful" authors having the highest likelihood of receiving replies and "unhelpful" authors the lowest.

The number of posts written by each author is selected at random, with "typical" authors usually posting fewer than ten times, with most posting only once or twice. Both "helpful" and "unhelpful" author post more frequently, with the number of posts of each of these authors selected uniformly between 20 and 200. Each post is assigned to one of four different forums. "Typical" authors contribute posts to only one forum per author, "helpful" authors post in one or two forums per author, and each "unhelpful" author may post in two to four forums. Each forum contains multiple threads which are created dynamically during the simulation.

Posts are added to threads sequentially. Each author is assigned a total number of posts and membership in a subset of the forums. For each new post, an author i is selected at random with probability proportional to the number of posts remaining for the author and a corresponding forum is selected. Author i either starts a new thread, or posts to an existing thread k, with probabilities 30 percent and 70 percent, respectively. An existing thread k is selected with probability proportional to the "reliability score" of the most reliable author in the thread. After the thread is selected, the author j to which author i is replying is selected, again with probability proportional to the "reliability score" of each author who previously posted in thread k. The reliability scores are 0.05, 0.6 and 1.0 for "unhelpful", "typical" and "unhelpful" authors, respectively; this means that "helpful" authors are significantly more likely to receive replies than "unhelpful" authors (see Figures 2a and 2b).



(a) Distribution of number of replies received per author for each author class.



(b) Distribution of number of replies posted per author for each author class.

Figure 2: Posting patterns of different author classes. "Helpful" authors both receive and post many replies, while "unhelpful" authors post many replies but receive relatively few replies. "Typical" authors have relatively few posts, so they neither receive or post many replies.

This process creates a set of threads with variable length and identifies each post either as the initial post in a new thread, or as a reply to a specific author. This allows us to create an accurate adjacency matrix for author interactions. This method generated a data set of 47495 posts in 14164 threads by 2000 unique authors.

The PageRank algorithm was applied to the author interaction data from this simulated set of forums. The results are plotted in Figure 3. We can see that the rescaled author interaction score allows us to clearly distinguish between the two classes of authors with many posts: "helpful" and "unhelpful" authors. While both classes have the same distribution in terms of total posts (Figure 4a), the "helpful" authors have higher author interaction rankings compared with the "unhelpful" authors, who tend to have relatively low interaction scores (Figure 4b). This demonstrates that when there are two classes of authors that post frequently, but

elicit different levels of response from other authors in the forum, the PageRank method for generating author interaction scores can allow us to distinguish between these classes so that only the posts of "unhelpful" authors can be discarded.

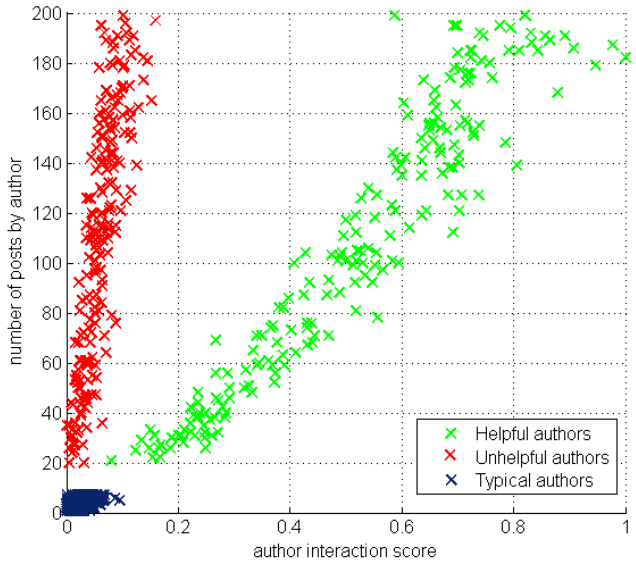
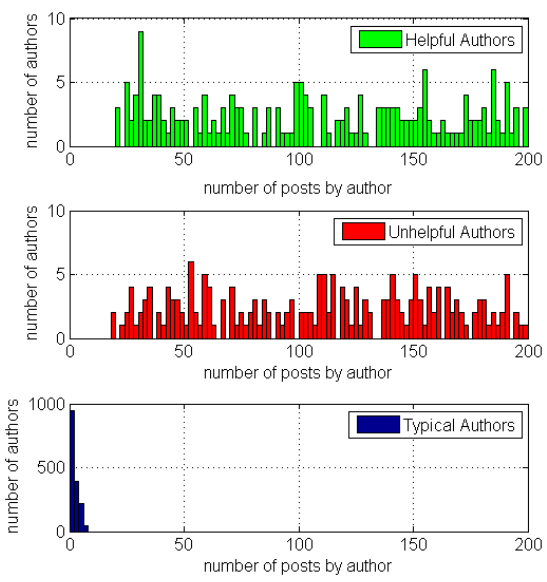
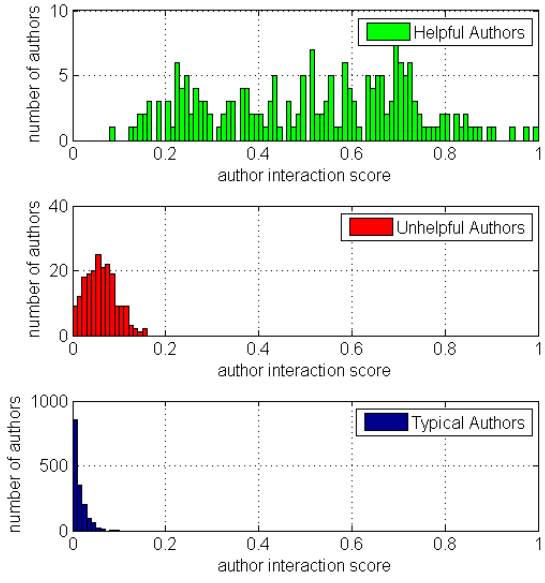


Figure 3: Plot of the author interaction score, rescaled to the interval $[0, 1]$, against the total number of posts by the author for simulated forum data. Each 'x' in the plot represents one author, with "helpful" authors in green,"unhelpful" authors in red and "typical" authors in blue. The three classes of authors can be divided into distinct clusters using the number of posts per author and author interaction score.



(a) Distribution of number of posts per author for each author class, corresponding to the vertical axis on the plot in Figure 3.



(b) Distribution of author interaction scores for each author class, corresponding to the horizontal axis on the plot in Figure 3.

Figure 4: Comparison of the posting patterns and author interaction scores for each author class. The distributions of number of posts per author are similar for "helpful" and "unhelpful" authors (4a), but the distributions for author interactions allow us to distinguish between these two classes (4b).

4.3 Burst Detection on VNS forum data

Our data set which we have worked on contains the documents collected from the self-reported forum posts regarding health outcomes and it also contains the terms used in the documents along with the document indices and the corresponding time points of their posting. The total number of documents including the time points of their posting is 28,279. We transformed the data to a term-document matrix which contains the frequency of each term in each document. We then applied the Non-Negative Matrix Factorization (NMF) on the term document matrix to cluster the documents into 25 different topics. We chose the number of classes as 25 based on some previous analysis of the data in SAS which tells us that roughly 25 topics are represented in the documents. This NMF clustering gave us 25 different clusters corresponding to 25 different topics where each cluster contains all the documents related to the particular topic and the corresponding dates when the documents were posted. We found a lot of missing dates and a lot of documents that were posted on exactly same date and we deleted them all to clean the data since this burst detection method does not work with repeated time points as the interval between consecutive messages is zero in that case. So we worked with roughly 6,000 unique time points. We applied our burst detection method on each topic at a time taking the time points of the forum posts related to that topic as the input of the algorithm. The output gives us the nested maximal time intervals which correspond to the bursts of different intensities. The output also generates the tree diagram which exhibits the hierarchical structure of the bursts of different intensities over the whole time line. The root of the tree obviously corresponds to the burst of lowest intensity that covers the whole time line of the forum posts of the particular topic.

Figure 5: Summary of Level Changes for 25 Topics

Topic	# of posts	Levels	Topic	# of posts	Levels
1	444	1, 2, 2, 3	14	139	1, 2, 2
2	541	1, 2 , 3	15	22	1
3	209	1, 2, 3, 4, 3	16	154	1, 2, 3, 4, 3
4	76	1, 2	17	435	1, 2, 3, 2
5	434	1, 2, 3	18	191	1, 2, 3, 4, 3
6	525	1, 2, 3, 3, 3	19	147	1,2
7	147	1, 2	20	260	1, 2, 3, 4, 3, 3
8	172	1, 2, 2	21	788	1,2,2
9	96	1, 2	22	104	1,2
10	106	1	23	333	1, 2, 3, 2
11	194	1, 2, 2	24	231	1, 2, 3, 3, 4, 3
12	105	1, 2	25	117	1,2,3
13	379	1, 2, 2, 2			
Total # of posts: 6,349					

The table in figure 5 has the number of documents for each topic that we have worked with and it also shows the bursts of different patterns and intensities for them. Higher level indicates higher intensity of bursts. To explain, topic 3 contains 209 unique time points corresponding to 209 posts and there are four different levels of bursts appearing in the pattern 1, 2, 3, 4, 3 which means the root of the tree corresponds to level 1 over the entire time line which contains one interval that is burst of intensity 2 and it contains two subintervals that are bursts of intensity 3 and the first of these 2 sub intervals of intensity 3 contains a further sub interval that is burst of intensity 4. Again topic 15 has only one interval which is the whole time line of level 1 which indicates there is no burst of the topic 15 at any point of time. We present the hierarchical structures of the bursts of four topics using rooted tree diagram below using four topics. To illustrate the meaning of the burst plots, we first demonstrate an example using topic 20, which is described by terms “patient”, “reported”, “physician”,

Figure 6: Time Intervals for Intensity Changes of Topic 20

Level	Start date (timestamp)	Start date	End date (timestamp)	End date
1	38058	3/12/2004	41313	2/8/2013
2	40343	6/14/2010	41215	11/2/2012
3	40628	3/26/2011	40749	7/25/2011
4	40654	4/21/2011	40742	7/18/2011
3	40867	11/20/2011	40888	12/11/2011
3	40967	2/28/2012	41051	5/22/2012

"indicated", "stimulation". For modeling purposes, all the dates are represented using "timestamps" available from the VNS dataset. Figure 6 shows the actual time interval corresponding to the timestamps in the SAS dataset. For a specific topic, the burst plots show the hierarchical structure of bursts of different intensities over time intervals.

Topic 20's tree structure clearly shows bursts of different intensities from 1 through 4 occurring at different points of time. The root time interval (38058, 41313) of level 1 contains an interval (40343, 41215) that is burst of intensity 2 which contains 3 sub intervals (40628, 40749), (40867, 40888) and (40967, 41051) all of which are bursts of intensity 3 and the first burst of intensity 3 contains a further sub interval (40654, 40742) that is burst of intensity 4. Topic 15 does not have any burst since there is no change of state at any point of time and hence there is just one level represented in the tree diagram. For topic 21, the tree has two levels. Level 1 contains 2 sub intervals that are bursts of intensity 2. For topic 16, the tree diagram has bursts of 4 different intensities. It is clear that more number of bursts of different intensities indicates frequent changes in the rate of arrivals of the posts along the time.

Topic 6 has 3 different level of bursts as we can see from the tree diagram. Interval (38111, 41331) contains one sub interval (40329, 41331) that is burst of level 2 which further contains 3 sub intervals that are bursts of intensity 3. We compare the burst plot with the histogram plot of the frequencies of the documents over the time intervals. We find that the time interval which shows burst of level 1 in the burst plot contains very few posts in the histogram, i.e. the arrival rate of posts in that interval is very slow. Again the interval indicating burst of level 2 has more posts in a short period of time meaning the rate of arrival of posts is higher than level 1 and the intervals showing burst of intensity 3 has relatively more posts indicating further increase in the rate of arrival of the posts in that time period. So we see that the method of burst detection matches well with the histogram plot of the relative frequencies. Also we give the table of normalized frequencies (frequencies divided by interval width) of the documents over the same time intervals as in the burst plot , which clearly indicates a well match between the two.

Using the results of burst detection we can assign each document with a number which is basically the intensity of the burst on the time interval that the document belongs to. If the document belongs to an interval which is a burst of intensity 2, then we assign the number 2 to that document and so on. So we can score all the documents related to each topic depending on the intensity of the bursts they experience. These level scores of the documents along with the corresponding time intervals can then be used to determine if the documents contain trending topics or not. If a burst with a high intensity stays for a longer time probably it gives indication of a trend in the topic that these documents correspond to. We can associate some weight to each burst and then compute some weighted measure for each topic and then rank the topics to find out which topic exhibits the most prominent rising and falling pattern over a limited period of time.

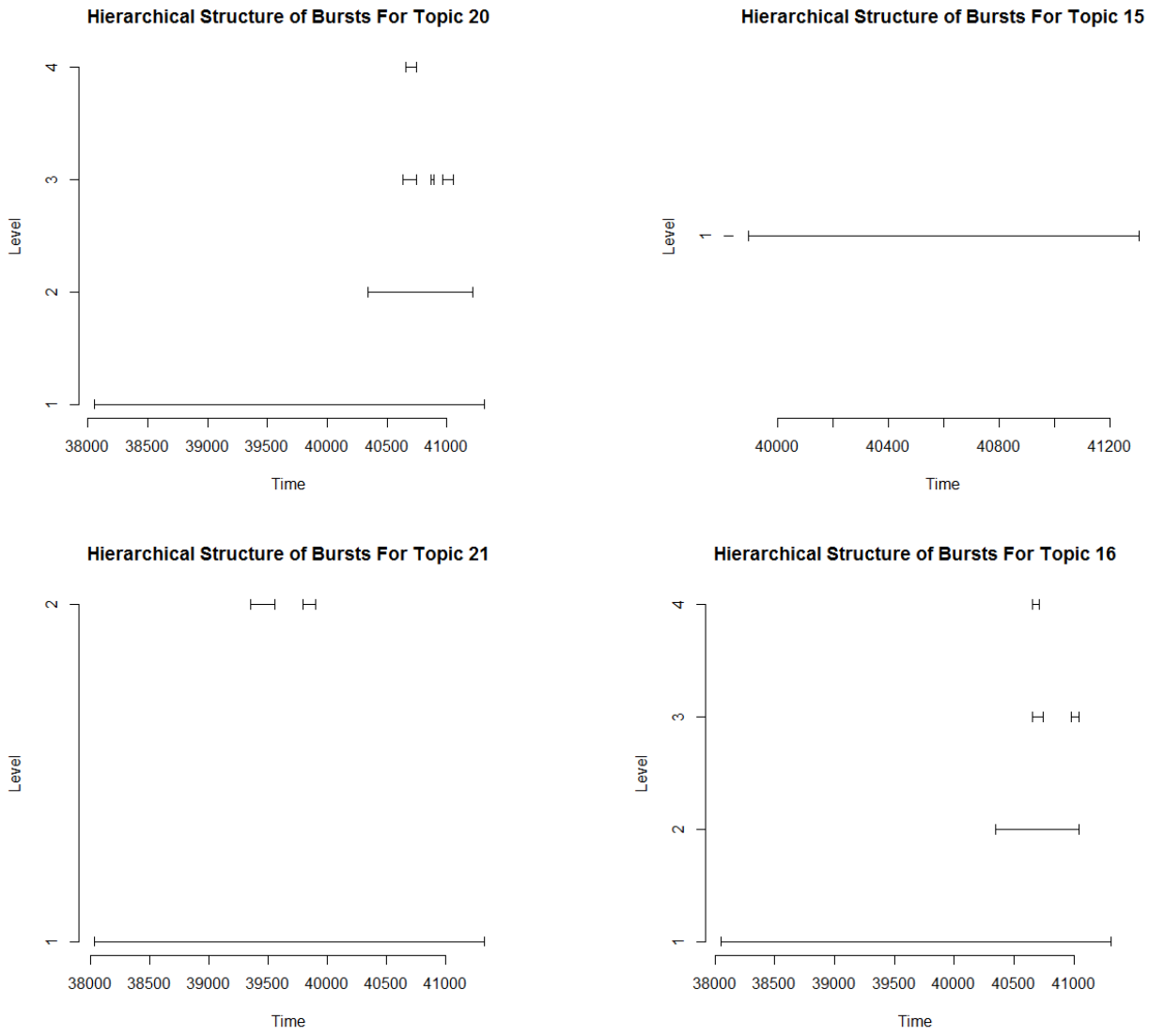
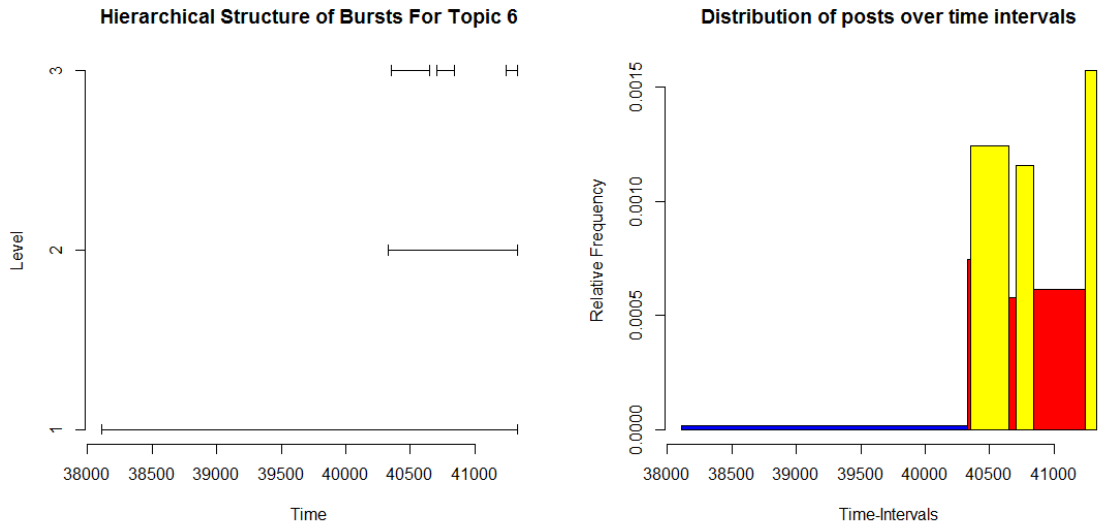


Figure 7: Burst Plots of Topic 20, 15, 21 and 16. Topic 20 is described by terms "patient", "reported", "physician", "indicated", "stimulation". Topic 15 is described by terms "date","information","event", "attempts", "number", "model", "additional", "type", "unsuccessful", "outcome", "reported", "patient". Topic 21 is described by terms "vas", "therapy", "implanted", "depression", and "problems". Topic 16 is described by terms "death", "event", "cause", "patient", "manufacturer", "system" and "passed".

Figure 8: Comparison between the burst plot for topic 6 and distribution of posts over different time intervals. Topic 6 is characterized by the following terms: "ia","ita","dona","youa","disorder" and "natasha".



(a) Normalized frequency by class width for topic 6

Time intervals	Frequency normalized by class width
38111, 40329	0.0077
40329, 40352	0.3913
40352, 40646	0.6531
40646, 40702	0.3036
40702, 40842	0.6071
40842, 41239	0.3224
41239, 41331	0.8261

5 Summary and Future Work

Web-based health forums contain a significant amount of information relevant to postmarking drug and medical device safety. Identifying which data posted on these media are useful and which are not is a challenging task. In this report we have proposed methods for generating features that together can be used to determine which documents and authors are producing useful content. We take advantage of the existing text mining features in SAS Text Miner, and propose new methods that provide additional insights. These new methods consider the sentiment or "emotions" within posts to detect possible sources of bias. We analyze author interactions to determine which authors are receiving many reponses from the forum community. Finally, we consider how topics discussed in the posts evolve over time by detecting bursts of posts about a particular topic.

The problem of determine the relevance of posts on web-forums is challenging and there are many areas for improvement. During our project we identified several challenges that arise when working with this type of data and application. First, unstructured data, especially spontaneous self-reported data from internet forums, is often incomplete. For example, although we proposed a method for ranking authors interactions based on the reply-structure of the forum posts, we currently to do not have access to data detailed enough to compute this on real web-data. Another challenge using data web-based sources is that the size of the datasets grow very quickly. This means that the methods we use need to be able to scale to handle large data sets. Annotation by domain experts is necessary to get good performance on supervised classification methods, however, it can be expensive because this is labor-intensive and tedious process.

There are many approaches we did not have time to explore during this project. One such approach is to group all the documents from a single thread and analyze this group of posts as a single "document" and record the length of thread and number of authors within the thread as additional features. This may be useful as a preprocessing step for identifying trending topics. If more detailed information can be collected about the reply-structure of the web forum we could do further analysis of not only individual authors within the forums, but also communities of authors.

References

[1] Iskowitz, M. "Social media provided unheard early warning on Avandia." Medical Marketing and Media, October 2010.

[2] Page, L. Brin, S. Motwani, R. & Winograd, T. (1999). "The PageRank citation ranking: Bringing order to the web."

[3] Brunet, J. P. Tamayo, P. Golub, T. R. & Mesirov, J. P. (2004). "Metagenes and molecular pattern discovery using matrix factorization." Proceedings of the National Academy of Sciences of the United States of America, 101(12), 4164–9.

[4] Park, H. (2012). "Nonnegative Matrix Factorization for Clustering Co-authors."

[5] Shahnaz, F. Berry, M. W. Pauca, V. P. & Plemmons, R. J. (2006). "Document clustering using nonnegative matrix factorization." Information Processing & Management, 42(2), 373–386.

[6] Xu, W. Liu, X. & Gong, Y. (2003). "Document Clustering Based On Non-negative Matrix Factorization." Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, p. 267–273.

[7] Gaujoux, R. et. al. (2010). "A flexible R package for nonnegative matrix factorization." BMC Bioinformatics 11.1, p. 367.

[8] Wolff, M. and Wallis, M. (2013). "Methods and Application for Determining the Integrity and Veracity of Medical Device Safety Related Data in Social Media." Pharmaceutical Industry SAS Users Group (PharmaSUG) Annual Conference.

[9] Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.

[10] Albright, R. (2004). "Taming Text with the SVD." SAS Institute Inc., Cary, North Carolina.

[11] Kleinberg, J. (2002), "Burst and Hierarchical Structure in Streams." Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.