

# Fifteenth Mathematical and Statistical Modeling Workshop for Graduate Students

19 – 29 July 2009  
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 at the Center for Research in Scientific Computation at North Carolina State University (NCSU) in Raleigh, North Carolina, 19 – 29 July 2009.

This was the fifteenth such workshop at NCSU. It brought together 39 graduate students from 33 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 a team of about 6 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 half-hour 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.ncsu.edu/crsc/events/imsm09/>

## **Sponsors**

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

## **Presenters**

Erik Gilleland, National Center for Atmospheric Research  
John Langstaff, Environmental Protection Agency  
Jordan Massad, Sandia National Laboratories  
Howard McLeod, UNC Institute for Pharmacogenomics  
Frank Meyer, Republic Mortgage Insurance Company  
John Peach, MIT Lincoln Laboratory

## **Faculty mentors**

Mansoor Haider, NCSU, Mathematics  
Elizabeth Mannshardt-Shamseldin, Duke, Statistics  
Alison Motsinger-Reif, NCSU, Statistics  
Brian Reich, NCSU, Statistics  
Jeff Scroggs, NCSU, Mathematics  
Ralph Smith, NCSU, Mathematics  
Richard Smith, UNC, Statistics

## **Student Participants**

Aaron Brown, Erin Byrne, Heejun Choi, Lili Ding, Robertas Gabrys, Chad Griep, Matthew Heaton, Magathi Jayaram, Matthias Katzfuss, Noory Kim, Jenya Kirshstein, Nitesh Kumar, Busaba Laungrungrong, Yi Li, Michael Lomuscio, Mario Morales, Glen Dale Pearson, Kathryn Pedings, Jacob Porter, William Proctor, Karthik Raghuram, Shahla Ramachandar, Gregory Richards,

Nishantha Samarakoon, Varada Sarovar, Maryam Shafahi, Chongyi Shen, Tyler Skorczewski, Kirk Soodhalter, Jie Sun, Eugene Vecharynski, Duy Vu, Min Wang, Jelani Wiltshire, Hongzia Yang, Jingyan Zhang, Peng Zhong, Kun Zhou, James Zou

## Projects

- **Modelling the effects of air pollution on public health**  
John Langstaff (EPA) and Brian Reich (NCSU Statistics)
- **Dosing predictions for the anticoagulant Warfarin**  
Michael Wagner (UNC School of Pharmacy) and Alison Motsinger-Reif (NCSU Statistics)
- **Severe Weather under a Changing Climate: Large Scale Indicators of Extreme Events**  
Erik Gilleland (National Center for Atmospheric Research), Elizabeth Mannshardt-Shamseldin (Duke Statistics and SAMSI) and Richard Smith (UNC Statistics)
- **Stress Tests, Toxic Assets, TARP, and Buying Your First Home**  
Frank Meyer (Republic Mortgage Insurance Company) and Jeff Scroggs (NCSU Mathematics)
- **Resource Issues Impacting National Security**  
John Peach (MIT Lincoln Laboratory) and Mansoor Haider (NCSU Mathematics)
- **High-Frequency, Low-Impact Switching of an RF MEMS Switch Without Pull-In Instability**  
Jordan Massad (Sandia National Laboratories) and Ralph Smith (NCSU Mathematics)

# Modeling the effects of air pollution on public health

Erin Byrne<sup>a</sup>, Heejun Choi<sup>b</sup>, Busaba Laungrungrong<sup>c</sup>, Michael Lomuscio<sup>d</sup>,  
Nishantha Samarakoon<sup>e</sup>, Jie (Rena) Sun<sup>f</sup>

*Problem Presenter:*

John Langstaff  
Environmental Protection Agency

*Faculty Mentor:*

Brian Reich  
North Carolina State University

## Abstract

We analyzed air pollution, exposure estimates, and mortality to compare relative rates of mortality associated with airborne particulate matter smaller than 10 microns ( $PM_{10}$ ) to those of the Air Pollution Exposure (APEX) simulated exposure estimates to  $PM_{10}$ . We estimated the form of the relationship between  $PM_{10}$  concentration and mortality and APEX exposure estimates and mortality. To estimate the relative risk associated with  $PM_{10}$  and estimated exposure for each city, we built generalized additive models that included adjustments for weather variables, long-term trends, and the presence of other pollutants. We completed a case study on data sets from Cook County, IL (Chicago) and Cuyahoga County, OH (Cleveland). A simulation study was conducted to compare the predictive abilities of the APEX estimates on mortality to those of ambient concentrations. Conclusions were discussed based on the results of the case study and simulation study.

## 1 Introduction and Motivation

Many recent studies [1] [2] [3] have investigated the association between air pollution and mortality. The US EPA sets standards as to the level of air quality that must be maintained [4]. Although many convincing trends have been observed, questions remain. A clearer picture of the effects of pollution on mortality would aid in the development of appropriate air quality standards.

Air pollution is composed of many gaseous pollutants and various types of particulate matter (PM) which are grouped according to the diameter of the particle. Although the mechanisms are still unknown, it is thought that PM of different diameters may affect the body in different ways, some more harmful than others. For instance, if PM is small enough it could be possible for it to penetrate deeper into the lungs than larger PM. This deeper penetration could result in a higher probability of a health response. In our study we consider PM of diameters less than 10 micrometers ( $PM_{10}$ ). This choice allows for incorporation of larger particles that could have an effect, while still incorporating the smallest particles into our dataset. The amount and size of PM is closely monitored by the EPA in each city. PM detecting sensors are placed throughout the cities to collect data and monitor the air quality.

Many former studies have been carried out by calculating the average density of pollutant in the air of a particular city and linking it to the levels of mortality within that city. However, this type of model makes unrealistic assumptions about the uniformity of pollutant and exposure. Although the average density of air pollutant can be a good indicator of the overall air quality of a city, the actual densities at any given locations can vary wildly. For instance, the density of PM would be much smaller in an office building with filtered air than at an industrial factory. Therefore, an individual in the office building would be exposed to much lower

---

<sup>a</sup>University of Colorado Boulder

<sup>b</sup>Purdue University

<sup>c</sup>Arizona State University

<sup>d</sup>Western Carolina University

<sup>e</sup>Kansas State University

<sup>f</sup>University of Michigan

densities of PM than an employee in the factory. Hence the actual exposure to PM by individuals in a city can be much different than average pollutant concentrations. It is the goal of this paper to consider a model of the effects of air pollution on mortality that incorporates individual exposure to the pollutant rather than simply the average ambient concentration.

In order to determine the amount of pollutant that individuals in a city are actually exposed to, the EPA has developed a stochastic human exposure simulator model. This model generates virtual individuals, follows them throughout their daily activities, and records their PM<sub>10</sub> exposure. This model was created by recording the day-to-day activities of actual people within a city. In this paper we will try to relate the simulated individual exposure to PM to deaths due to cardiopulmonary, cardiovascular, and respiratory disease.

The paper proceeds as follows. Section 2 describes the data used in our analysis, as well as the treatment of missing data. Section 3 describes the model we used to carry out the case studies of the data sets from Chicago and Cleveland. Section 4 describes the simulation study we carried out to investigate the potential benefits of including individual exposures in a health response model. Section 5 reports the results of several models run for the case studies of the two cities. Section 6 discusses conclusions and possible future work based on our findings.

## 2 Data

The data come from two sources. Mortality, meteorological data, and ambient air pollution data were obtained from the National Mortality, Morbidity, and Air Pollution Study (NMMAPS) database and are described in Section 2.1. The simulated personal exposures using the Air Pollution Exposure (APEX) simulator, described in Section 2.2. The treatment of missing data is described in Section 2.3.

### 2.1 NMMAPS Data

The NMMAPS data are daily time series of mortality, meteorological data, and air pollutants for Cook County, Illinois (Chicago) and Cuyahoga County, Ohio (Cleveland) taken from the NMMAPS data set completed in 2003. The observations were collected over a 14-year period from 1987 to 2000 for a total 5114 daily observations for each city. These data are freely available through the Internet-based Health & Air Pollution Surveillance System at the Johns Hopkins Bloomberg School Of Public Health <sup>1</sup>.

The mortality data considered are daily counts of deaths derived from death certificates and consist of county residents who died of various non-accidental causes. Counts are broken down into three age categories (under 65, 65-74, and 75 and over), and separate counts are also included for various causes of death (respiratory failure, cardiovascular disease, and chronic obstructive pulmonary disease). The mean daily non-accidental deaths is 115 for Chicago, ranging from a minimum of 69 deaths to maximum of 411. For Cleveland, the mean is 37 deaths ranging from a minimum of 17 to a maximum of 68. Deaths among people age 75 and older made up approximately half of the total deaths for each county.

Daily meteorological data included are mean temperature (°F), dew point temperature (°F), and the average of relative humidity (%). Air pollution data include the maximum hourly recorded measurement averaged over all stations in the county (PM10max) and the mean daily value over the entire county (PM10mean). Air pollutant concentrations measured in ppb are ozone (O<sub>3</sub>), sulfur dioxide (SO<sub>2</sub>), nitrogen dioxide (NO<sub>2</sub>), and carbon monoxide (CO). PM<sub>10</sub> and PM<sub>2.5</sub> concentrations are measured in  $\mu\text{g}/\text{m}^3$ .

### 2.2 APEX Data

The Air Pollution Exposure model (APEX) is a stochastic, multipollutant model designed to simulate population exposure to air pollutants. It is applied to specific study areas (in this case, Cook County, IL and Cuyahoga County, OH) and uses census data, such as gender and age, to generate demographic characteristics for various simulated individuals. It then constructs an activity diary to represent the sequence of activities and micro-environments each simulated individual experiences over a given period. Air quality of each micro-environment is determined from user-specified methods and, when coupled with breathing-rate and physiological parameters, generate the individual's exposure to particular air pollutants. The data input to

---

<sup>1</sup><http://www.ihapss.jhsph.edu/>

the APEX model are briefly described in this section and more information can be found in the APEX User’s Guide and Technical Support Document [6].

Our study considers only exposure levels to PM<sub>10</sub>, and the exposure modeling is based on PM<sub>10</sub> concentrations measured at ambient air monitors in the areas being modeled. These data include the daily and hourly concentration measurements from the monitoring data maintained in EPA’s Air Quality System <sup>2</sup>. Daily maximum temperatures input to APEX come from the NMMAPS data set.

The population demographics are input to APEX for every Census tract in the study areas. 1,344 tracts in Chicago (Cook County) and 502 tracts in Cleveland (Cuyahoga County) were modeled. Population counts and employment probabilities by age and gender are used to develop representative profiles of hypothetical individuals for the simulation. Tract-level population counts by age in one-year increments, from birth to 99 years, come from the 2000 Census of Population and Housing Summary File 1 <sup>3</sup>. The employment data is described on the Census web site <sup>4</sup>. In addition to using estimates of employment by tract, APEX also incorporates home-to-work commuting data. Commuting data for all pairs of tracts were derived from the 2000 Census and were collected as part of the Census Transportation Planning Package (CTTP).

To ensure that individuals’ daily activities are reasonably represented within APEX, it is important to integrate working patterns into the assessment. The APEX commuting data were collected as part of the CTTP, and contain tabulations by place of residence, place of work, and the flows between the residence and work. These data are available from the U.S. Department of Transportation, Bureau of Transportation Statistics <sup>5</sup>.

Daily activity patterns for individuals in a study area are obtained from detailed diaries that are compiled in the Consolidated Human Activity Database (CHAD) [5]. The time-location-activity diaries input to APEX contain information regarding an individual’s age, gender, race, employment status, occupation, day-of-week, daily maximum hourly average temperature, the location, start time, duration, and type of each activity performed. Much of this information is used to best match the activity diary with the generated personal profile, using age, gender, employment status, day of week, and temperature as first-order characteristics. The approach is designed to capture the important attributes contributing to an individual’s behavior.

The APEX model produces estimates of exposures in  $\mu\text{g}/\text{m}^3$  corresponding to the concentration of PM<sub>10</sub> a simulated individual is exposed to in a given day according to their activities. The sample of individual exposures is then analyzed to find the mean, maximum, and percentiles of the population exposure distribution. For example, the covariate “P95” is the concentration of PM<sub>10</sub> a person in the 95th percentile of the exposure distribution experiences. The APEX exposure concentrations are lower than the ambient concentrations of PM<sub>10</sub> in the atmosphere, as shown in Figure 1.

### 2.3 Imputation of Missing Data

Our data set has a large amount of missing values for multiple covariates for various reasons. As our primary objective is to analyze the association between PM<sub>10</sub> concentration, APEX and the number of deaths, we dismissed the observations that do not have PM<sub>10</sub> measurements. The dismissed PM<sub>10</sub> values appeared to be missing at random and included approximately 7.8% of the observations. Also, APEX data are only available from January 1, 1994 forward, so only these dates are included. After these two considerations, 2557 daily observations for each city remain.

The average relative humidity information was absent in all observations past January 1, 1998, for both Chicago and Cleveland. Alternate data were obtained through National Climate Data Center (NCDC) website to help complete the data set. Daily minimum and maximum relative humidities were recorded for the majority of days during the 14 year span of observations considered, and the average of the minimum and maximum was computed from these observations to provide an additional measurement. Multiple months were not recorded, but the new data filled in enough of the missing time period to allow analysis to continue. Due to the difference in available measurements from the two sources, the NCDC maximum, minimum, and derived average relative humidity measurements replaced the NMMAPS average relative humidity in our analysis.

<sup>2</sup><http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm>

<sup>3</sup><http://www.census.gov/prod/cen2000/doc/sf1.pdf>

<sup>4</sup><http://www.census.gov/population/www/cen2000/phc-t28.html>

<sup>5</sup><http://transtats.bts.gov>

Chicago July 1st 1994 to July 31st 1994

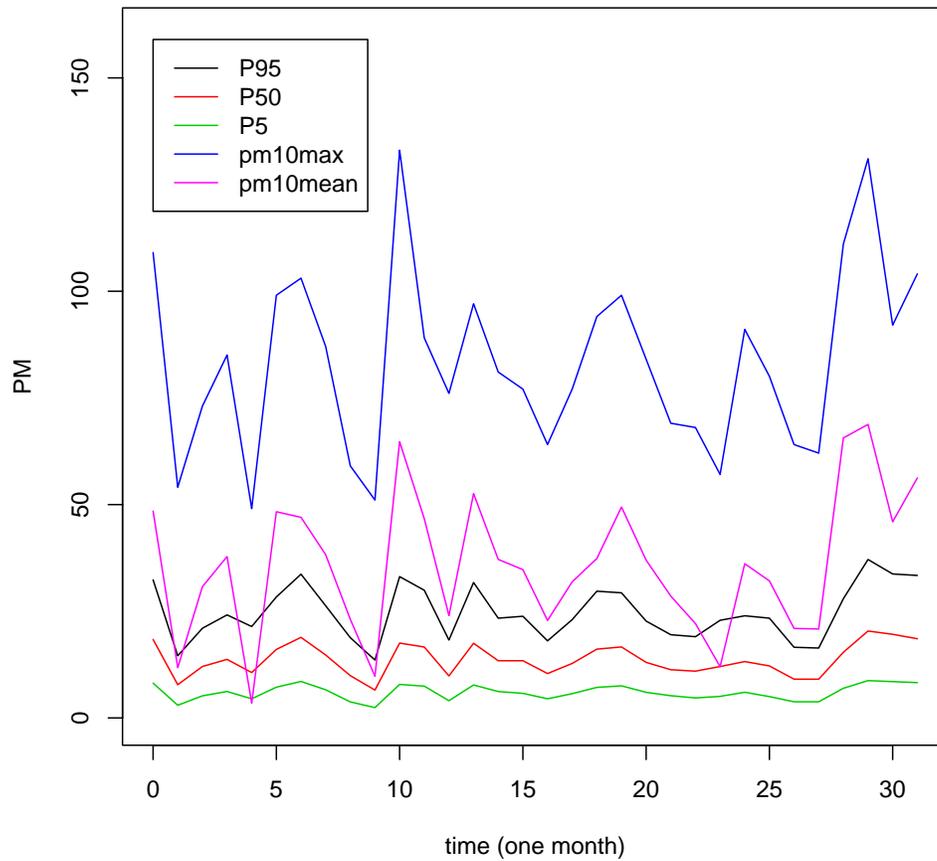
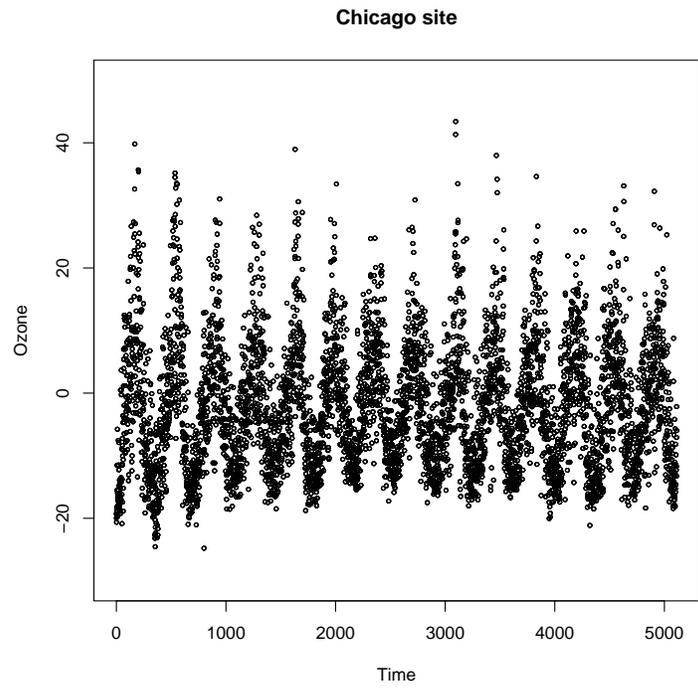
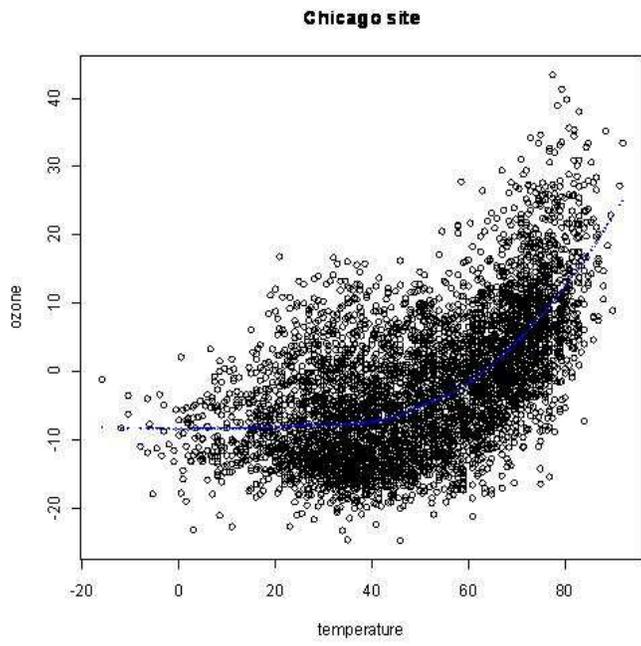


Figure 1: Ambient  $PM_{10}$  concentrations and APEX  $PM_{10}$  exposure concentration, Chicago. PM measured in  $\mu g/m^3$ . P95, P50, and P5 are the concentration of  $PM_{10}$  a person in the 95th, 50th, and 5th percentile of the population exposure distribution experiences, respectively. pm10max and pm10mean are the maximum and mean daily  $PM_{10}$  concentrations.

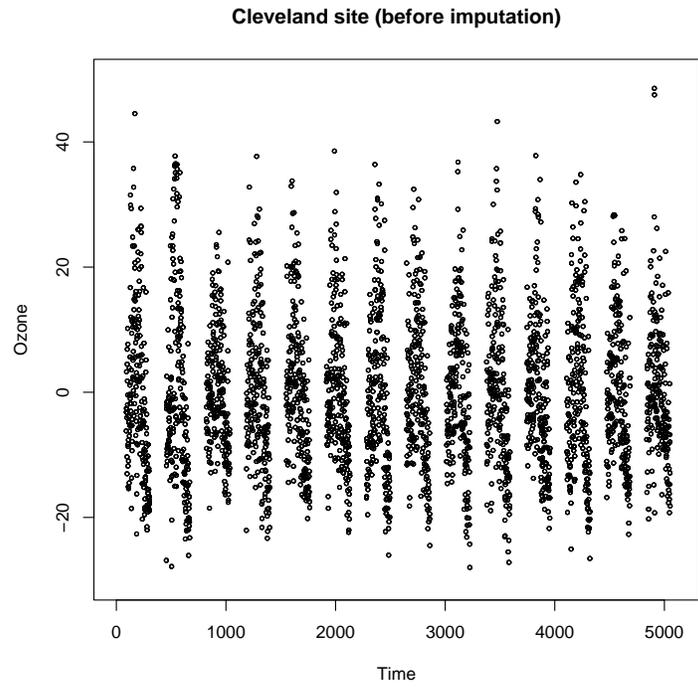


(a) Time Series for Ozone Concentration

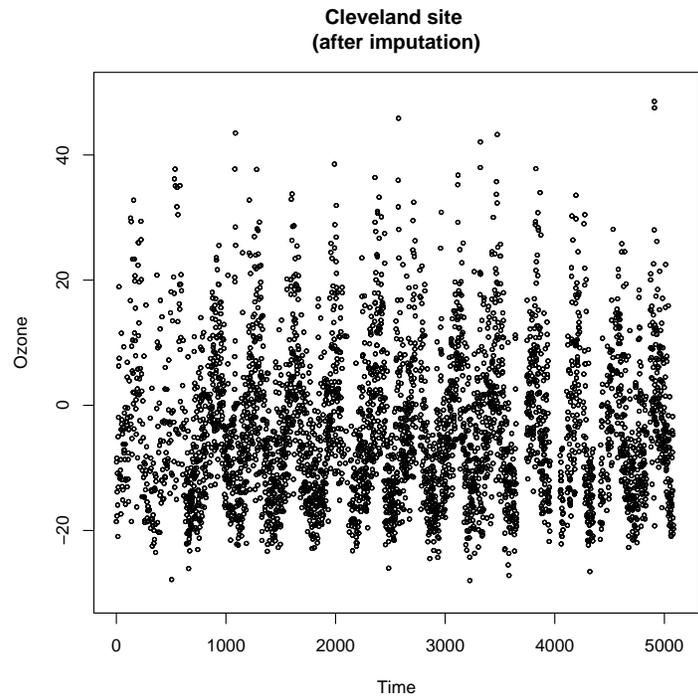


(b) Temperature vs. Ozone for Chicago

Figure 2: Ozone Concentrations for Chicago.



(a) Original Time Series



(b) Time Series with Imputed Data

Figure 3: Time Series of Ozone Concentrations for Cleveland.

Predictors	Primary reasons for inclusion
Day of the week	To allow for different baseline mortality rates within each day of the week
Time	To adjust for long-term trends and seasonality
Temperature Dewpoint Relative humidity	To control for the known effects of weather on mortality
Mean CO Mean Ozone Mean SO <sub>2</sub> Mean NO <sub>2</sub>	To control the potential effects of pollutants on mortality

Table 1: Potential confounding factors in the model

The ozone measurement had systematic missing observations for Cleveland because the city does not monitor ozone levels between November 1 and March 31 each year. We imputed the data from other covariates known to affect ozone concentrations. Because Chicago has a near-complete data set for ozone levels (Figure 2a), it was used to explore general trends between ozone concentration and various meteorological data. A linear regression model appeared appropriate with predictors of maximum relative humidity, average daily temperature, day of the week, daily precipitation, total sunshine, and seasonality. A quadratic term for temperature was included to account for a marked increase in ozone for temperatures over 50 (Figure 2b). Seasonality was accounted for by including a sinusoidal term for time as a predictor. Model fit was checked and the model explained most of the variation in ozone measurement. Missing ozone data for Cleveland was filled in based on this model (Figure 3).

Several other covariates were still missing a small number of randomly distributed observations and were chosen to be deleted. PM<sub>2.5</sub> was not collected until March 1998, so we did not consider PM<sub>2.5</sub> in our analysis due to its absence in the majority of the observations and its inclusion in the atmospheric measurement of PM<sub>10</sub>. After all omissions, there were 2535 observations for Chicago and 2542 observations for Cleveland.

### 3 Model Description

Generalized additive models (GAM) [7] were used in our analysis to explore the relationship between the count of deaths as dependent variable, and additive predictors through a nonlinear link function. A log-linear GAM was adopted in this case for the Poisson response – the number of events. Day of the week, temperature, dewpoint, relative humidity, seasonality, mean CO level, mean Ozone level, mean SO<sub>2</sub> level and mean NO<sub>2</sub> level were included in the model as potential confounding effects (Table 1). To be able to get an overall picture of the air pollution effect on the mortality, we modeled the cardiovascular deaths, respiratory deaths, all cause deaths excluding accident, and chronic obstructive pulmonary disease separately.

The outcome variable,  $Y_t$ , is defined as the total number of events of a given type (e.g. cardiovascular death) at time  $t$ . It is assumed to follow a Poisson distribution with mean  $\mu_t$ , and  $Y_t$  are independent of given  $\mu_t$ . The GAM model has the form

$$\log(\mu_t) = S(X_t) + \beta Z_t,$$

where  $S(X_t) = S(X_{1t}) + S(X_{2t}) + \dots + S(X_{nt})$ ,  $Z = (Z_{1t}, Z_{2t}, \dots, Z_{mt})^T$  and  $\beta = (\beta_1, \dots, \beta_m)$ . The covariates that are assumed to have linear impact on the log form of the count data are included in  $Z$ , while the others are included in  $X$ . The non-linear effects of  $X$  are modeled using separate spline functions with various number of knots.

The lag effects of PM<sub>10</sub> on the number of deaths were also analyzed. For example, the effect of PM<sub>10</sub> level from the previous day on the mortality of current day was evaluated, referred as ‘one-day lag effect’. In our analysis, no lag, one-day lag, two-day lag, and three-day lag of PM<sub>10</sub> were fitted in the log linear GAM model separately to compare the model difference. Akaike’s Information Criterion (AIC), coefficient estimate, and  $p$ -value were used as the model selection criterion. AIC, a measure of the goodness of fit for an estimated statistical model, is defined as  $AIC = 2k - 2 \log(L)$ , where  $L$  is the maximized value of the likelihood function

for the estimated model and  $k$  is the effective degrees of freedom being estimated, to keep the balance between the goodness of fit and model complexity.

Statistical software R 2.9 was used to carry out the analysis. Package MGCV was used to analyze the Poisson generalized additive models.

## 4 Simulation Study

In this section, we conduct a simulation study to investigate the potential benefits of accounting for individual exposure in an epidemiological study of air pollution and a health response. We simulate data to closely resemble the Chicago data. The idea of simulation is to generate pollution and response data at the individual level using the APEX distribution and a logistic model. We randomly sample  $x_{it}$  for the  $i^{th}$  person on the  $t^{th}$  day from the data generated by APEX model (every 5th percentile of estimated PM<sub>10</sub> exposure varying from P5 to P95) and then  $y_{it}$  as the response based on these predictors. Consider the logistic regression model for a response variable  $y_{it} \sim \text{Bernoulli}(\pi_{it})$  where  $\pi_{it}$  is the probability of death on the  $t^{th}$  day in the city as described below.

$$E(y_{it}) = P(y_{it} = 1|x_{it}) = \pi(x_{it}) = \frac{\exp(\beta_{0i} + \beta_{1i}x_{it})}{1 + \exp(\beta_{0i} + \beta_{1i}x_{it})} \rightarrow \mathbf{Logit}(\pi(x_{it})) = \log\left(\frac{\pi(x_{it})}{1 - \pi(x_{it})}\right) = \beta_{0i} + \beta_{1i}x_{it}$$

$\beta_{0i} \sim N(\mu_0, \sigma_0^2)$  and  $\beta_{1i} \sim N(\mu_1, \sigma_1^2)$ . This random-coefficient model allows for heterogeneity in the effect of PM on different individuals in the population.

We report the simulation study in which there are  $d = 120$  days and a population size in the city of  $n = 10000$  for  $s = 100$  data sets.

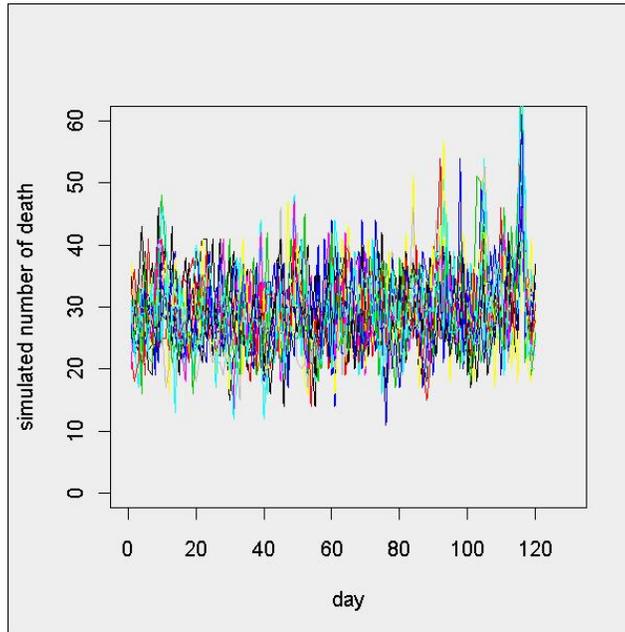


Figure 4: Plot of simulated PM10 exposure data.

In the next step, we generated  $n=10000$  Bernoulli trials,  $y_{it}$  with the probability of  $\pi(x_{it})$ . The simulated response variable for a particular day is the sum of above Bernoulli values, i.e.  $y_t = \sum_{i=1}^n y_{it}$ . So, we studied the properties of a Poisson random variable regarding our simulated data  $y_t$ . Figure 4 shows a sample data set.

	PM10Mean	PM10Max	P50	P95	Mean
Model 1: $\beta_0 \sim N(-7, 0.01)$ $\beta_1 \sim N(0, 0)$	9	7	6	6	6
Model 2: $\beta_0 \sim N(-7, 0.01)$ $\beta_1 \sim N(0.0001, 0.01)$	18	19	22	20	22
Model 3: $\beta_0 \sim N(-7, 0.01)$ $\beta_1 \sim N(0.01, 0.01)$	73	73	82	79	81
Model 4: $\beta_0 \sim N(-7, 0.01)$ $\beta_1 \sim N(0.1, 0.01)$	100	100	100	100	100
Model 5: $\beta_0 \sim N(-7, 0.01)$ $\beta_1 \sim N(0.5, 0.01)$	100	100	100	100	100

Table 2: Estimated power for the GLM models.

All the simulations were conducted in *R* using the GLM to fit the model. We fit the Poisson regression models taking the simulated  $y_t$  as the response variable and PM10mean, PM10max, P50, P95, and mean exposure as the predictor. Selecting one variable at a time we find the  $p$ -value of including this predictor in the model. Table 2 reports the proportion of  $p$ -values  $< 0.05$  for each method, i.e. the power of the method.

There are two things that we are interested in this study Type I error and the power of the test. Type I error is defined by the probability that we reject the null hypothesis if the  $p$ -value  $< 0.05$  when the true value of is  $\beta_1 = 0$ . Type I error is explored through the first model (where  $\beta_1 = 0$ ). The small percentage of rejecting the null hypothesis indicates the small chance to include the predictor variable in the model when there is truly no  $\beta_1$ . It can be seen that all of these parameters provide a 6 - 9% chance of rejecting the null hypothesis, slightly greater than the acceptable range ( $p$ -value = 0.05). For the power of the test, we look at the probability that we reject the null hypothesis when  $\beta_1 \neq 0$ . The power of the test is examined by models 2 - 5. The high percentage of rejecting the null hypothesis represents the ability to include the predictor variable in the model. Overall, these five variables have a high chance ( $> 70\%$ ) to add to the model when it should truly be in the model. The APEX exposure estimates (P50, P95, and mean) are better predictors in all four models (2–5) than the ambient concentrations (PM10mean and PM10max) since they generally have a higher percentage of rejecting the null hypothesis when  $\beta_1 \neq 0$ .

## 5 Results

To determine the appropriate lag model, we considered lag for zero to three days and ranked them according to their AIC value, and whether or not the lag was statistically significant. Our findings were that one day lag was the best model for Chicago, and same day PM<sub>10</sub> Cleveland. Because of this we chose to model both cities separately, Chicago with one day lag and Cleveland with no lag. These choices of lag within each city were consistent across all causes of death in each city respectively.

For each city we ran several models, shown in Tables 3 and 4. In both cities we varied our responses to consider different types of death. We also varied the indicator of PM<sub>10</sub>, including the mean level of PM<sub>10</sub> for a given day (pm10mean), the maximum of the hourly measurements of PM<sub>10</sub> (pm10max), 50th percentile of the population exposure distribution (P50), the 95th percentile of the population exposure distribution (P95), and the mean of the population exposure distribution (mean).

The predictor that is most commonly considered in epidemiological studies is pm10mean, however we can see from the Table 3 that for the Chicago data set pm10max has smaller AIC for all four of our responses. One plausible explanation for pm10max having a stronger association than pm10mean could be that the death response is triggered only after PM<sub>10</sub> has reached some threshold level. In this case the pm10max would be a

	COD	AIC	RR	Log RR	P-Values
PM10 Mean	Resp	13021.9	1.01	0.011	0.0099
PM10 Max	Resp	13018.6	1.01	0.011	0.0022
P50	Resp	13023.1	1.04	0.037	0.0179
P95	Resp	13021.5	1.02	0.021	0.0086
Mean	Resp	13022.6	1.04	0.037	0.0140
PM10 Mean	CVD	18199.4	1.005	0.005	0.0216
PM10 Max	CVD	18195.1	1.01	0.005	0.0045
P50	CVD	18203.3	1.01	0.013	0.0999
P95	CVD	18200.2	1.01	0.009	0.0295
Mean	CVD	18202.4	1.01	0.014	0.0691
PM10 Mean	COP	10761	1.01	0.011	0.0675
PM10 Max	COP	10759.9	1.01	0.011	0.0394
P50	COP	10762.6	1.03	0.031	0.1719
P95	COP	10762	1.02	0.018	0.1146
Mean	COP	10762	1.03	0.030	0.1506
PM10 Mean	NAD	20401.6	1.006	0.006	< 0.0001
PM10 Max	NAD	20401.3	1.005	0.005	< 0.0001
P50	NAD	20409	1.018	0.018	0.0005
P95	NAD	20402.8	1.011	0.011	0.0001
Mean	NAD	20407	1.018	0.018	0.0003

Table 3: Chicago. The relative risk, RR, correlates to a  $10^{\frac{\mu g}{m^3}}$  increase in  $PM_{10}$ .  $RR = e^{10\beta}$ . NAD = Non-accidental death. Mean level of  $PM_{10}$  for a given day (pm10mean), the maximum of the hourly measurements of  $PM_{10}$  (pm10max), 50th percentile of the population exposure distribution (P50), the 95th percentile of the population exposure distribution (P95), and the mean of the population exposure distribution (mean).

good predictor of whether that threshold was reached.

Although pm10max is the best predictor for Chicago, the goal of our study is to determine how well the APEX data predicts each response. From Table 3 we see that P95 is a statistically significant predictor for respiratory death and non-accidental death, and that each model has an AIC value comparable to that of the pm10max predictor. However, P95 does not appear to be a good predictor of COPD or CVD.

For Cleveland, Table 4, P95 is a noticeably better predictor of respiratory and non-accidental death. For the Chicago data the difference between AIC values for the pm10max and P95 models is less than 3 for the respiratory response and less than 1.5 for non-accidental death response, however for the Cleveland data we find that the P95 model has an AIC value 2 less than the pm10max model for the respiratory response and about 5.7 less than the pm10max model for the non-accidental death response. Hence P95 is a better predictor of non-accidental death in Cleveland, and at least as good of a predictor for respiratory death in Chicago and Cleveland, and non-accidental death in Chicago.

Upon analysis we discovered that the principal confounders that are significant to our models are mean temperature, time trend (which accounts for seasonal variation), relative humidity, and in our Chicago models carbon monoxide mean. Table 5 shows the  $p$ -values for each of the confounders for eight different models. As we saw above, pm10max and P95 models for respiratory and non-accidental death responses are the most interesting. Hence we only included these eight.

We can see from the table that mean temperature, date, and relative humidity are usually statistically relevant predictors of the respiratory and non-accidental death responses. In Chicago carbon monoxide mean is also a statistically significant predictor.

## 6 Conclusions

We studied the air pollution and its relationship to the mortality data in 2 selected cities, Chicago and Cleveland. The exposures and ambient are used to compare for fitting the model in each city. One problem we

	COD	AIC	RR	Log RR	P-Values
PM10 Mean	Resp	7605.7	1.026	0.026	0.0441
PM10 Max	Resp	7606.7	1.015	0.015	0.0848
P50	Resp	7602.1	1.094	0.090	0.0063
P95	Resp	7604.6	1.037	0.036	0.0232
Mean	Resp	7602.6	1.085	0.082	0.0079
PM10 Mean	CVD	11776.1	1.019	0.019	0.0001
PM10 Max	CVD	11778.4	1.012	0.012	0.0005
P50	CVD	11785.6	1.027	0.027	0.0443
P95	CVD	11783.2	1.016	0.016	0.0112
Mean	CVD	11784.8	1.027	0.027	0.0276
PM10 Mean	COP	6286.7	1.031	0.031	0.0663
PM10 Max	COP	6287.2	1.02	0.020	0.0918
P50	COP	6282	1.105	0.100	0.0226
P95	COP	6289	1.041	0.040	0.0603
Mean	COP	6282.7	1.092	0.088	0.0323
PM10 Mean	NAD	13461.1	1.015	0.015	< 0.0001
PM10 Max	NAD	13464.7	1.009	0.009	0.0001
P50	NAD	13463.3	1.037	0.036	< 0.0001
P95	NAD	13459	1.019	0.019	< 0.0001
Mean	NAD	13461.7	1.036	0.035	< 0.0001

Table 4: Cleveland. The relative risk, RR, correlates to a  $10 \frac{\mu g}{m^3}$  increase in  $PM_{10}$ .  $RR = e^{10\beta}$ . NAD = Non-accidental death. Mean level of  $PM_{10}$  for a given day (pm10mean), the maximum of the hourly measurements of  $PM_{10}$  (pm10max), 50th percentile of the population exposure distribution (P50), the 95th percentile of the population exposure distribution (P95), and the mean of the population exposure distribution (mean).

found is that there are many missing data, especially in Cleveland data set. The linear regression technique is applied to impute the missing data in Cleveland. In the other hand, we removed the missing data in Chicago as only a few are missing. The results show that the best predictor variables that will be included in the model vary by city to city. The max, ambient, and concentration variables are significant to the model for Chicago data whereas 95th percentile of exposure is only one significant variable for modeling Cleveland. We also investigated an accounting for individual variable by simulation method. The non-accidental deaths are generated by assuming Bernoulli distribution with the probability that was randomly selected from 5th 95th percentile of the exposures of the APEX model. Then, we fitted the data to the Poisson regression model and calculated the proportion that we reject the null hypothesis if p-value  $\leq 0.05$  when the null hypothesis is true. The simulation results present that the exposure can lead to a 5-10% increase in the power of test.

Several issues concern about the limitations in this study, for example, the small data set, and the variables assumed in the simulation model. First we picked only Chicago and Cleveland to fit the model and did the simulation for Chicago data. It can be extended to additional cities and we also recommend investigating these additional data as well as we did for Chicago. Second, we examine one variable in the simulation at a time which is the percentile of the  $PM_{10}$  exposure estimated by APEX model. It is interesting to add one or more variable to the model and also the higher order terms including the polynomial and interaction for the further study. Another issue to concern is that the results of simulation technique rely on the generated data only from the APEX model. It is possible to improve the data by combining the simulated exposures with the actual data.

	Chicago				Cleveland			
	pm10max		P95		pm10max		P95	
	resp	NAD	Resp	NAD	Resp	NAD	Resp	NAD
Mean Temperture	2.62e-08	2e-16	1.06e-08	2e-16	0.07	0.039	0.0683	0.0124
Date	7.54e-10	2e-16	7.89e-10	2e-16	0.04	0.0003	0.0431	1.05e-07
Relative humidity	0.0001	1.88e-06	0.0002	0.0001	0.007	0.0158	0.0272	0.0472
COmean	0.0124	0.009	0.013	0.0097	0.38	0.5131	0.4425	0.8706

Table 5: Table of  $p$ -values for each confounder for each model.

## References

- [1] Dominici, F., Daniels, M., Zeger, S.L., Samet, J.M. (2002). "Air Pollution and Mortality: Estimating Regional and National Dose-Response Relationships." *J. Amer. Statistical Assoc.* 97(457):100-111.
- [2] Daniels, M., Dominici, F., Samet, J. M., and Zeger, S. L. (2000), "Estimating PM10-Mortality Dose-Response Curves and Threshold Levels: An Analysis of Daily Time-Series for the 20 Largest U.S. Cities," *American Journal of Epidemiology*, 152:397-412.
- [3] Schwartz, J. (1994), "Air Pollution and Daily Mortality: A Review and Meta Analysis," *Environmental Research*, 64:36-52.
- [4] U.S. Environmental Protection Agency (1996), "Review of the National Ambient Air Quality Standards for Particulate Matters: Policy Assessment of Scientific and Technical Information," OAQPS Staff paper.
- [5] McCurdy, T., Glen, G., Smith, L., Lakkadi, Y. (2000). "The National Exposure Research Laboratory's Consolidated Human Activity Database. *J. Exposure Anal. Environ. Epidemiol.* 10:566-578.
- [6] U.S. Environmental Protection Agency (2008). Total Risk Integrated Methodology (TRIM) - Air Pollutants Exposure Model Documentation (TRIM.Expo / APEX, Version 4) Users Guide and Technical Support Document. U.S. Environmental Protection Agency, Research Triangle Park, NC. EPA-452/B-08-001a,b. [http://www.epa.gov/ttn/fera/human\\_apex.html](http://www.epa.gov/ttn/fera/human_apex.html).
- [7] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.
- [8] Haneuse, S., and Wakefield, J. (2008). "The combination of ecological and case-control data". *J.R. Statist. Soc. B*, 70:73-93.

# Dosing Predictions for the Anticoagulant Warfarin

Group Members: Lili Ding\*, Noory Kim†, Mario Morales‡,  
Dale Pearson§, Jacob Porter¶, Maryam Shafahi||

Advisors: Alison Motsinger-Reif\*\*, Micheal Wagner††

July 28, 2009

---

\*University of Cincinnati, Cincinnati, OH

†Towson University, Towson, MD

‡Hunter College, New York, NY

§Texas Tech University, Lubbock, TX

¶University of California, Davis

||University of California, Riverside

\*\*North Carolina State, Raleigh, NC

††University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

## Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Methods and Results . . . . .	3
1.2	Data Cleaning . . . . .	3
1.2.1	Subgrouping . . . . .	4
1.2.2	Data Summary . . . . .	4
1.2.3	Remarks . . . . .	8
<b>2</b>	<b>Tests for Demographic Differences Between Subgroups</b>	<b>9</b>
<b>3</b>	<b>Measure of performance of the NEJM algorithm by subgroups</b>	<b>15</b>
<b>4</b>	<b>Dosing Predictions for the Anticoagulant Warfarin</b>	<b>16</b>
4.1	Univariate analysis . . . . .	16
4.2	Whole Population And Subpopulation Models . . . . .	20
<b>5</b>	<b>Summary</b>	<b>23</b>
<b>6</b>	<b>Acknowledgments</b>	<b>24</b>
<b>7</b>	<b>References</b>	<b>24</b>

## List of Figures

1	published NEJM algorithm . . . . .	15
2	Frequency of different CYP2C9 variants in OMB population groups . . . . .	19
3	Univariate analysis of CYP2C9 for different subpopulations . . . . .	19
4	Univariate analysis of different gene variants for the whole population . . . . .	20
5	Model performance by race . . . . .	23

## List of Tables

1	Continuous Variables, mean(SD): OMB Categories . . . . .	5
2	Continuous Variables, mean(SD): Race Reported Categories (count $\geq$ 50) . . . . .	5
3	Demographic Variables: OMB Categories . . . . .	5
4	Demographic Variables: Race Reported Categories (count $\geq$ 50) . . . . .	6
5	Genotype Counts: OMB Categories . . . . .	7
6	Genotype Counts: Race Reported Categories (count $\geq$ 50) . . . . .	8
7	Contingency Analysis for OMB Races . . . . .	10
8	Contingency Analysis for the Koreans . . . . .	11
9	Contingency Analysis for the Chinese . . . . .	12
10	Contingency Analysis for the Japanese . . . . .	13
11	Contingency Analysis for the Malay and Mixed . . . . .	14
12	NEJM algorithm [6] Performance . . . . .	16
13	Univariate Analysis for OMB Races . . . . .	17
14	Univariate Analysis for the Asian Subgroups . . . . .	18
15	Multivariable Analyses . . . . .	22
16	Model performance . . . . .	23

# 1 Introduction

Since its discovery and approval for human use over fifty years ago, Warfarin has become the most widely used oral anticoagulant worldwide, prescribed over 20 million times each year in the United States. It is commonly used to prevent abnormal blood clotting.

Despite its widespread use it ranks among the top ten drugs in the United States associated with serious adverse events, as doses required for safe and effective treatment can vary by a factor of 10 among individuals. [NEJM] Thus traditional methods of establishing an appropriate warfarin dosage level based on clinically measurable data end up using a significant degree of trial and error. This can put patients at risk from life threatening side effects from over-dosage, which can result in hemorrhaging of blood, or under-dosage, which can result in lack of efficacy in preventing those medical conditions for which warfarin is prescribed (such as stroke).

The efficacy of warfarin for an individual is gauged by the standard measure for blood clotting ability is the International Normalized Ratio (INR). The standard target INR range is between 2.0 and 3.0. For some patients achieving a stable level within this target may not occur even after 90 days.

Currently the process for achieving a stable dose for a patient requires often months of a trial and error process, decreasing efficacy and increasing risk of side effect during that time. By improving on the dosing process, the amount of time it take to reach stable dose, and the corresponding risks to patients could be reduced. It has been previously reported that the use of genetic markers within a predictive dosing algorithm could reduce this time.[2, 3, 4, 7]

Many studies have reliably and consistently found that variations in two genes, cytochrome P4502C9 (CYP2C9) and Vitamin K epoxide reductase complex subunit 1 (VKORC1), are significantly correlated with warfarin dose, and the U.S. Food and Drug Administration (FDA) includes this finding on the drug label. <sup>1</sup> [5]

Recently, the International Warfarin Pharmacogenetics Consortium (IWPC), composed of 21 research groups from around the world, grouped their studies to assemble a collection of data from 5700 patients and derived a global dosing algorithm based on 5052 of those patients. [6] This pharmacogenomic algorithm had two genetic and six clinical predictors, and predicted over 43% of the variation in therapeutic dosing within the study, outperforming a dosing algorithm with only clinical predictors (26%)and a traditional fixed dose approach (0%).

In the current study, we use the data currently available from the IWPC consortium to evaluate several aspects of the performance of the published algorithm. Additionally, we explore the potential of refined dosing algorithms for different racial and ethnic subpopulations in the data, as unique risk factors and genetic backgrounds may change the optimal model within these groups, with the goal of minimizing error in the models for each individual.

## 1.1 Methods and Results

## 1.2 Data Cleaning

The current dataset contained and had clinical and genetic information from 5933 patients. We only considered those subjects who successfully reached a stable dose of warfarin, and whose target INR was between 2

---

<sup>1</sup>That these two genes play a part in the regulation of blood clotting levels has been known since around the early 1990s. The genes encode proteins involved in the regulation of vitamin K, which is essential for blood clotting. The VKORC1 is part of the Vitamin K epoxide reductase complex which activates vitamin K, thereby increasing blood clotting ability. CYP2C9 is resident in the liver and its metabolism of warfarin deactivates the effect of warfarin on blood clotting.

This connection between warfarin and vitamin K can explain some of the instability in the dosage of warfarin. A shift in vitamin K levels (through ingestion of spinach, for example) can throw off the relationship between warfarin dosage and INR.

and 3, to be consistent with the approach of the IWPC study, resulting in a total of 5106 individuals available for analysis. An additional 3 patients were excluded from analysis based on inconsistent or unique racial and ethnic identifiers (“H”, “Black British”, “Black African”), resulting in a total of 5103 patients for analysis. The data had already been cleaned substantially to increase consistency among data entries, and we did some further data cleaning. For example, we eliminated some differences in usage of capital letters and hyphens (“other” vs. “Other”) as well as some variations in terminology (“African-American”, “African American”, “Black or African-American” were all changed to “Black”). Additionally, due to deviations from normality, the outcome variable of interest (the therapeutic dose of warfarin), was transformed using a square root transformation.

### 1.2.1 Subgrouping

The data set had multiple data fields pertaining to race and ethnicity, including “race reported” and “OMB”. We relied primarily on these two variables to generate ethnic and racial subsets of the table. Individuals that appeared to belong to more than one racial category (having been listed under reported race as “Intermediate” or “Other Mixed Race”) were grouped together under “Other”. Not all racial and ethnic groups had sample sizes deemed large enough to justify separate analysis. For example, only 42 patients (0.008%) were classified as “Hispanic”.<sup>2</sup> Only groups with a minimum of 50 individuals per group were used for stratifying the datasets into groups.

### 1.2.2 Data Summary

Data was evaluated for the following variables:

- Continuous variables: height, weight, therapeutic warfarin dose, square root of therapeutic dose, target INR, INR reported with stable warfarin dose
- Categorical variables: clinical
  - gender, age
  - project site, reason for warfarin treatment
  - other health conditions:
    - diabetes, congestive heart failure and or valve replacement
  - other medications:
    - amiodarone, enzyme inducer status(carbamazepine, phenytoin, OR rifampin), aspirin, acetaminophen, simvastatin, atorvastatin (Lipitor), fluvastatin, lovastatin, pravastatin, rosuvastatin, cerivastatin, sulfonamide antibiotics, macrolide antibiotics, anti-fungal azoles, herbal medications or vitamin supplements
- Categorical variables: genetic CYP2C9, VKORC1 SNPs (1639, 497, 1173, 1542, 3730, 2255, 4451)

Tables 1 through 6 display summaries of the majority of variables included in the current dataset. Summary means, standard deviations, proportions, and total counts are included as appropriate for each variable.

---

<sup>2</sup>These 42 patients were classified as “Hispanic” under the current US Office of Management and Budget (OMB) racial and ethnic classification scheme (enacted in 1997).

Variable	Whole Data	OMB White	OMB Black	OMB Asian	OMB Other
Count	5103	3035	663	958	447
Height (cm)	168.74 (10.29)	171.03 (9.94)	170.44 (10.42)	161.49 (7.74)	166.21 (9.57)
Weight (kg)	80.62 (22.09)	84.65 (19.92)	92.36 (27.07)	61.61 (11.74)	76.58 (19.76)
Therapeutic Dose	32.90 (17.36)	33.73 (17.83)	41.69 (19.5)	24.59 (11.12)	32.06 (13.72)
Warfarine Dose sqrt	5.56 (1.43)	5.63 (1.44)	6.28 (1.49)	4.84 (1.07)	5.53 (1.20)
Target INR	2.52 (0.18)	2.54 (0.15)	2.52 (0.13)	2.39 (0.17)	2.66 (0.23)
Reported INR	2.43 (0.33)	2.46 (0.32)	2.44 (0.33)	2.32 (0.32)	2.50 (0.37)

Table 1: Continuous Variables, mean(SD): OMB Categories

Variable	Chinese	Japanese	Korean	Malay	Mixed	NA
Count	335	227	259	82	80	464
Height (cm)	162.08 (7.78)	161.46 (8.29)	160.61 (7.58)	161.57 (6.8)	162.88 (9.79)	169.73 (9.61)
Weight (kg)	62.28 (12.91)	61.41 (10.41)	59.62 (9.87)	63.69 (13.33)	70.17 (14.17)	87.09 (21.33)
Therapeutic Dose	22.81 (10.51)	19.99 (9.03)	28.4 (8.33)	25.69 (9.7)	32.31 (10.93)	32.69 (14.68)
Warfarine Dose sqrt	4.66 (1.03)	4.36 (0.97)	5.27 (0.78)	4.97 (0.98)	5.6 (0.98)	5.58 (1.24)
Target INR	2.5 (0)	2.31 (0.25)	2.25 (0)	2.5 (0)	2.77 (0.25)	2.66 (0.12)
Reported INR	2.37 (0.36)	2.31 (0.25)	2.21 (0.3)	2.48 (0.26)	2.61 (0.36)	2.53 (0.34)

Table 2: Continuous Variables, mean(SD): Race Reported Categories (count  $\geq$  50)

Variable		Whole Data	OMB White	OMB Black	OMB Asian	OMB Other
Count		5103	3035	663	958	447
Gender	female	2234 (43.8%)	1183 (39%)	371 (56%)	448 (46.8%)	232 (51.9%)
	male	2866 (56.2%)	1852 (61%)	292 (44%)	507 (52.9%)	215 (48.1%)
	Total Known	5100 (99.9%)	3035 (100%)	663 (100%)	955 (99.7%)	447 (100%)
Age (years)	10-19	13 (0.3%)	6 (0.2%)	2 (0.3%)	2 (0.2%)	3 (0.7%)
	20-29	119 (2.3%)	69 (2.3%)	21 (3.2%)	14 (1.5%)	0 (0%)
	30-39	230 (4.5%)	103 (3.4%)	49 (7.4%)	51 (5.3%)	15 (3.4%)
	40-49	545 (10.7%)	231 (7.6%)	109 (16.4%)	147 (15.3%)	27 (6%)
	50-59	1000 (19.6%)	505 (16.6%)	163 (24.6%)	234 (24.4%)	0 (0%)
	60-69	1211 (23.7%)	692 (22.8%)	151 (22.8%)	276 (28.8%)	58 (13%)
	70-79	1341 (26.3%)	940 (31%)	120 (18.1%)	185 (19.3%)	98 (21.9%)
	80-89	597 (11.7%)	465 (15.3%)	42 (6.3%)	34 (3.5%)	92 (20.6%)
	90+	33 (0.6%)	24 (0.8%)	6 (0.9%)	1 (0.1%)	96 (21.5%)
	Total Known	5089 (99.7%)	3035 (100%)	663 (100%)	944 (98.5%)	56 (12.5%)

Table 3: Demographic Variables: OMB Categories

Variable		Chinese	Japanese	Korean	Malay	Mixed	NA
Count		335	227	259	82	80	464
Gender	female	142 (42.4%)	63 (27.8%)	176 (68%)	41 (50%)	51 (63.8%)	191 (41.2%)
	male	193 (57.6%)	161 (70.9%)	83 (32%)	41 (50%)	29 (36.3%)	273 (58.8%)
	Total	335 (100%)	224 (98.7%)	259 (100%)	82 (100%)	80 (100%)	464 (100%)
Age	10-19	1 (0.3%)	0 (0%)	1 (0.4%)	0 (0%)	2 (2.5%)	0 (0%)
	20-29	4 (1.2%)	0 (0%)	5 (1.9%)	3 (3.7%)	5 (6.3%)	1 (0.2%)
	30-39	21 (6.3%)	2 (0.9%)	19 (7.3%)	3 (3.7%)	8 (10%)	1 (0.2%)
	40-49	46 (13.7%)	8 (3.5%)	58 (22.4%)	26 (31.7%)	15 (18.8%)	21 (4.5%)
	50-59	79 (23.6%)	39 (17.2%)	89 (34.4%)	18 (22%)	26 (32.5%)	49 (10.6%)
	60-69	91 (27.2%)	73 (32.2%)	74 (28.6%)	24 (29.3%)	18 (22.5%)	109 (23.5%)
	70-79	81 (24.2%)	76 (33.5%)	11 (4.2%)	4 (4.9%)	5 (6.3%)	172 (37.1%)
	80-89	12 (3.6%)	14 (6.2%)	2 (0.8%)	4 (4.9%)	1 (1.3%)	106 (22.8%)
	90+	0 (0%)	1 (0.4%)	0 (0%)	0 (0%)	0 (0%)	5 (1.1%)
Age	Total	335 (100%)	213 (93.8%)	259 (100%)	82 (100%)	80 (100%)	464 (100%)

Table 4: Demographic Variables: Race Reported Categories (count  $\geq 50$ )

Genetic Marker (Count)	Genotype	Whole Data 5103	OMB White 3035	OMB Black 663	OMB Asian 958	OMB Other 447
CYP2C9	*1/*1	3637 (71.3%)	1907 (62.8%)	577 (87.0%)	835 (87.2%)	318 (71.1%)
	*1/*2	738 (14.5%)	619 (20.4%)	38 (5.7%)	1 (0.1%)	80 (17.9%)
	*1/*3	467 (9.2%)	336 (11.1%)	23 (3.5%)	75 (7.8%)	33 (7.4%)
	*2/*2	55 (1.1%)	49 (1.6%)	0 (0%)	0 (0%)	6 (1.3%)
	*2/*3	68 (1.3%)	63 (2.1%)	1 (0.2%)	0 (0%)	4 (0.9%)
	*3/*3	19 (0.4%)	15 (0.5%)	0 (0%)	1 (0.1%)	3 (0.7%)
	Total Known	4984 (97.7%)	2989 (98.5%)	639 (96.4%)	912 (95.2%)	444 (99.3%)
VKORC1 497	G/G	78 (1.5%)	74 (2.4%)	3 (0.5%)	0 (0%)	1 (0.2%)
	G/T	509 (10.0%)	464 (15.3%)	27 (4.1%)	6 (0.6%)	12 (2.7%)
	T/T	1281 (25.1%)	637 (21%)	360 (54.3%)	245 (25.6%)	39 (8.7%)
	Total Known	1868 (36.6%)	1175 (38.7%)	390 (58.8%)	251 (26.2%)	52 (11.6%)
VKORC1 1173	C/C	928 (18.2%)	455 (15%)	337 (50.8%)	46 (4.8%)	90 (20.1%)
	C/T	881 (17.3%)	556 (18.3%)	67 (10.1%)	162 (16.9%)	96 (21.5%)
	T/T	975 (19.1%)	207 (6.8%)	7 (1.1%)	732 (76.4%)	29 (6.5%)
	Total Known	2784 (54.6%)	1218 (40.1%)	411 (62.0%)	940 (98.1%)	215 (48.1%)
VKORC1 1542	C/C	794 (15.6%)	231 (7.6%)	26 (3.9%)	525 (54.8%)	12 (2.7%)
	C/G	1086 (21.3%)	806 (26.6%)	110 (16.6%)	138 (14.4%)	32 (7.2%)
	G/G	887 (17.4%)	645 (21.3%)	169 (25.5%)	44 (4.6%)	29 (6.5%)
	Total Known	2767 (54.2%)	1682 (55.4%)	305 (46%)	707 (73.8%)	73 (16.3%)
VKORC1 1639	A/A	954 (18.7%)	315 (10.4%)	12 (1.8%)	595 (62.1%)	32 (7.2%)
	A/G	1364 (26.7%)	1063 (35.0%)	91 (13.7%)	98 (10.2%)	112 (25.1%)
	G/G	1352 (26.5%)	800 (26.4%)	431 (65%)	13 (1.4%)	108 (24.2%)
	Total Known	3670 (71.9%)	2178 (71.8%)	534 (80.5%)	706 (73.7%)	252 (56.4%)
VKORC1 2255	C/C	448 (8.8%)	306 (10.1%)	80 (12.1%)	46 (4.8%)	16 (3.6%)
	C/T	551 (10.8%)	371 (12.2%)	28 (4.2%)	135 (14.1%)	17 (3.8%)
	T/T	657 (12.9%)	116 (3.8%)	10 (1.5%)	524 (54.7%)	7 (1.6%)
	Total Known	1656 (32.5%)	793 (26.1%)	118 (17.8%)	705 (73.6%)	40 (8.9%)
VKORC1 3730	A/A	313 (6.1%)	183 (6%)	82 (12.4%)	38 (4%)	10 (2.2%)
	A/G	1011 (19.8%)	614 (20.2%)	203 (30.6%)	165 (17.2%)	29 (6.5%)
	G/G	1320 (25.9%)	465 (15.3%)	102 (15.4%)	734 (76.6%)	19 (4.3%)
	Total Known	2644 (51.8%)	1262 (41.6%)	387 (58.4%)	937 (97.8%)	58 (13%)
	NA	2459 (48.2%)	1773 (58.4%)	276 (41.6%)	21 (2.2%)	389 (87%)
VKORC1 4451	A/A	72 (1.4%)	67 (2.2%)	3 (0.5%)	1 (0.1%)	1 (0.2%)
	A/C	271 (5.3%)	227 (7.5%)	33 (5%)	4 (0.4%)	7 (1.6%)
	C/C	594 (11.6%)	246 (8.1%)	243 (36.7%)	81 (8.5%)	24 (5.4%)
	Total Known	937 (18.4%)	540 (17.8%)	279 (42.1%)	86 (9%)	32 (7.2%)

Table 5: Genotype Counts: OMB Categories

Genetic Marker (Count)	Genotype	Chinese 335	Japanese 227	Korean 259	Malay 82	Mixed 80	NA 464
CYP2C9	*1/*1	285 (85.1%)	218 (96%)	231 (89.2%)	63 (76.8%)	63 (78.8%)	315 (67.9%)
	*1/*2	0 (0%)	0 (0%)	0 (0%)	0 (0%)	11 (13.8%)	90 (19.4%)
	*1/*3	24 (7.2%)	9 (4%)	28 (10.8%)	7 (8.5%)	5 (6.3%)	47 (10.1%)
	*2/*2	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1.3%)	2 (0.4%)
	*2/*3	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	7 (1.5%)
	*3/*3	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (0.4%)
	Total Known	309 (92.2%)	227 (100%)	259 (100%)	70 (85.4%)	80 (100%)	463 (99.8%)
VKORC1 497	G/G	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	4 (0.9%)
	G/T	3 (0.9%)	0 (0%)	0 (0%)	0 (0%)	3 (3.8%)	7 (1.5%)
	T/T	200 (59.7%)	0 (0%)	26 (10%)	5 (6.1%)	20 (25%)	17 (3.7%)
	Total Known	203 (60.6%)	0 (0%)	26 (10%)	5 (6.1%)	23 (28.8%)	28 (6%)
VKORC1 1173	C/C	9 (2.7%)	0 (0%)	0 (0%)	7 (8.5%)	5 (6.3%)	83 (17.9%)
	C/T	59 (17.6%)	26 (11.5%)	31 (12%)	36 (43.9%)	2 (2.5%)	91 (19.6%)
	T/T	251 (74.9%)	201 (88.5%)	228 (88%)	39 (47.6%)	0 (0%)	27 (5.8%)
	Total Known	319 (95.2%)	227 (100%)	259 (100%)	82 (100%)	7 (8.8%)	201 (43.3%)
VKORC1 1542	C/C	250 (74.6%)	201 (88.5%)	23 (8.9%)	37 (45.1%)	3 (3.8%)	33 (7.1%)
	C/G	62 (18.5%)	26 (11.5%)	3 (1.2%)	38 (46.3%)	11 (13.8%)	151 (32.5%)
	G/G	8 (2.4%)	0 (0%)	0 (0%)	7 (8.5%)	9 (11.3%)	107 (23.1%)
	Total Known	320 (95.5%)	227 (100%)	26 (10%)	82 (100%)	23 (28.8%)	291 (62.7%)
VKORC1 1639	A/A	162 (48.4%)	201 (88.5%)	219 (84.6%)	2 (2.4%)	8 (10%)	4 (0.9%)
	A/G	36 (10.7%)	26 (11.5%)	30 (11.6%)	0 (0%)	43 (53.8%)	10 (2.2%)
	G/G	7 (2.1%)	0 (0%)	0 (0%)	3 (3.7%)	29 (36.3%)	15 (3.2%)
	Total Known	205 (61.2%)	227 (100%)	249 (96.1%)	5 (6.1%)	80 (100%)	29 (6.3%)
VKORC1 2255	C/C	9 (2.7%)	0 (0%)	0 (0%)	7 (8.5%)	3 (3.8%)	14 (3%)
	C/T	61 (18.2%)	26 (11.5%)	3 (1.2%)	38 (46.3%)	4 (5%)	10 (2.2%)
	T/T	250 (74.6%)	201 (88.5%)	23 (8.9%)	37 (45.1%)	0 (0%)	4 (0.9%)
	Total Known	320 (95.5%)	227 (100%)	26 (10%)	82 (100%)	7 (8.8%)	28 (6%)
VKORC1 3730	A/A	6 (1.8%)	0 (0%)	31 (12%)	7 (8.5%)	4 (5%)	3 (0.6%)
	A/G	62 (18.5%)	26 (11.5%)	0 (0%)	36 (43.9%)	13 (16.3%)	17 (3.7%)
	G/G	252 (75.2%)	201 (88.5%)	226 (87.3%)	39 (47.6%)	6 (7.5%)	8 (1.7%)
	Total Known	320 (95.5%)	227 (100%)	257 (99.2%)	82 (100%)	23 (28.8%)	28 (6%)
VKORC1 4451	A/A	1 (0.3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (0.6%)
	A/C	4 (1.2%)	0 (0%)	0 (0%)	0 (0%)	1 (1.3%)	15 (3.2%)
	C/C	33 (9.9%)	0 (0%)	26 (10%)	5 (6.1%)	5 (6.3%)	10 (2.2%)
	Total Known	38 (11.3%)	0 (0%)	26 (10%)	5 (6.1%)	6 (7.5%)	28 (6%)

Table 6: Genotype Counts: Race Reported Categories (count  $\geq 50$ )

### 1.2.3 Remarks

Despite data cleaning, there are many caveats in order that are the result of a lack of consistency among groups in (1) selecting patients and in (2) selecting data variables. An example of bias in selecting patients appears in the Korean study, whose 258 patients were all listed as having had a heart valve replacement. The location of consortium members could be partly responsible for the high number of white patients and the low number of Hispanic and African patients. Lack of uniformity in selecting data variables resulted in missing data. Despite a consensus within the IWPC on which variables to include in the overall study, only a few variables had information for all patients in the study.

The therapeutic mean dose for the OMB Asian subpopulation is significantly lower than for other subpopulations. Non-genetic factors such as differences in mean weight and target dosage are significantly lower as well, so these could account for at least some of the difference. This lack of consistency makes it difficult to assess the degree that genetic differences have a role in this difference.

While the uniformity of the data collected from the IWPC study is far from ideal, the lessons learned can

be incorporated could be valuable should a follow up study of this scale be conducted.

## 2 Tests for Demographic Differences Between Subgroups

To investigate potential clinical and genetic differences between the racial and ethnic subpopulations evaluated, we performed tests of associations for these demographic variables. All analyses were performed with JMP Version 7.0 ([www.jmp.com](http://www.jmp.com)) and SAS Version 9.1.3 ([www.sas.com](http://www.sas.com)).

For each racial and Asian subgroup, we compared all of the variables in the published NEJM algorithm [6]. In addition, we compared gender, age, indication for Warfarin treatment, Carbamazepine Tegretol, Phenytoin Dilantin, Rifampin or Rifampicin, Diabetes, Congestive Heart Failure, Valve Replacements, Aspirin, Acetaminophen or Paracetamol, Simvastatin Zocor, Atorvastatin Lipitor, Fluvastatin Lescol, Lovastatin Mevacor, Pravastatin Pravachol, Rosuvastatin Crestor, Cervastatin Baycol, Sulfonamide Antibiotics, Macrolide Antibiotics, Anti Fungal Azoles and Herbal Medications Vitamins.

Association analysis was performed in JMP software ([www.jmp.com](http://www.jmp.com)). Contrasts between subpopulation were performed with either Chi-square tests of association for categorical demographic variables, or Student's T tests for continuous outcomes (or nonparametric versions in the case of assumption violations). Two-tailed p-values for obtained for each test. For each test our null hypothesis assumed that there was not a correlation between the row variable and the column variable, and we used a conventional alpha rate of 0.05 as our cut-off for determining statistical significance. Finally, we used Microsoft Excel to plot our p-values for each subgroup.

In tables 7 - 11 every variable that has a p-value of less than .05 is denoted with an asterisk (\*). Our tables visualize differences between subgroups, and show that there is a significant difference between most of the variables for each subgroup. For example, the Chinese have a significant difference for VKORC1 than the other Asian subgroups as seen in the table. These significant associations suggest potentially important demographic differences that could be accounted for by the subpopulation specific models.

Variable	White vs. Black	White vs. Asian	White vs. Other	Asian vs. Other	Asian vs. Black	Black vs. Other
Gender	*	*			*	*
Age	*	*			*	
Ind. for Warfarin Treatment	*	*		*	*	
Amiodarone Cordarone	*	*	*	*	*	*
Carbamazepine Tegretol	*	*			*	*
Phenytoin Dilantin	*	*			*	*
Rifampin or Rifampicin	*	*			*	*
Enzyme Inducer Status		*			*	
CYP2C9 *1/*2	*	*		*	*	
CYP2C9 *1/*3	*	*			*	
CYP2C9 *2/*2	*	*				
CYP2C9 *2/*3	*	*				
CYP2C9 *3/*3						
CYP2C9_NA	*	*				
VKORC1.1639_NA	*				*	
VKORC1.1639_A/A	*	*		*	*	*
VKORC1.1639_A/G	*	*		*	*	
VKORC1.497_G/T	*	*		*	*	*
VKORC1.497_G/G	*	*		*	*	*
VKORC1.497_NA	*	*	*	*	*	
VKORC1.1173_C/C	*	*		*	*	
VKORC1.1173_C/T	*				*	*
VKORC1.1173_NA	*	*	*	*	*	
VKORC1.1542_C/C	*	*	*		*	*
VKORC1.1542_C/G	*	*				
VKORC1.1542_NA	*	*			*	*
VKORC1.3730_A/A	*	*		*	*	
VKORC1.3730_A/G	*	*			*	
VKORC1.3730_NA	*	*	*	*	*	
VKORC1.2255_C/C		*		*	*	
VKORC1.2255_C/T	*				*	*
VKORC1.2255_NA	*	*	*		*	*
VKORC1.4451_A/A	*	*				
VKORC1.4451_A/C	*	*		*	*	
VKORC1.4451_NA	*	*	*	*	*	
Diabetes	*	*	*		*	*
Congestive Heart Failure	*	*	*	*	*	*
Valve Replacement	*	*		*	*	
Aspirin	*	*	*	*	*	
Acetaminophen	*	*		*	*	
Simvastatin Zocor	*	*		*	*	
Atorvastatin Lipitor	*	*	*	*	*	*
Fluvastatin Lescol	*				*	*
Lovastatin Mevacor	*	*			*	*
Pravastatin Pravachol	*	*	*	*	*	*
Rosuvastatin Crestor	*	*			*	*
Cervastatin Baycol	*	*			*	*
Sulfonamide Antibiotics	*	*			*	*
Macrolide Antibiotics	*	*			*	*
Anti Fungal Azoles	*	*			*	*
Herbal Medications	*	*			*	

Table 7: Contingency Analysis for OMB Races

Variable	Korean vs. Chinese	Korean vs. Japanese	Korean vs. Malay	Korean vs. Mixed	Korean vs. NA
Gender	*	*	*		*
Age	*	*			*
Ind. for Warfarin Treatment	*	*	*	*	*
Amiodarone Cordarone	*	*	*	*	
Carbamazepine Tegretol	*	*	*	*	*
Phenytoin Dilantin	*	*	*	*	*
Rifampin or Rifampicin	*	*	*	*	*
CYP2C9 *1/*2				*	*
CYP2C9 *1/*3		*			
CYP2C9 *2/*3					*
CYP2C9_NA	*		*		
VKORC1.1639_NA	*	*	*		*
VKORC1.1639_A/A	*		*	*	*
VKORC1.1639_A/G			*	*	*
VKORC1.497_G/T				*	*
VKORC1.497_NA	*	*		*	*
VKORC1.1173_C/C	*		*	*	*
VKORC1.1173_C/T			*	*	*
VKORC1.1173_NA	*			*	*
VKORC1.1542_C/C	*	*	*		
VKORC1.1542_C/G	*	*	*	*	*
VKORC1.1542_NA	*	*	*	*	*
VKORC1.3730_A/A	*		*	*	
VKORC1.3730_A/G	*		*		*
VKORC1.3730_NA	*			*	*
VKORC1.2255_C/C	*		*	*	*
VKORC1.2255_C/T	*	*	*	*	
VKORC1.2255_NA	*	*	*		*
VKORC1.4451_A/C					*
VKORC1.4451_NA		*			*
Diabetes	*	*	*	*	*
Congestive Heart Failure	*	*	*	*	*
Valve Replacement	*	*	*	*	*
Aspirin	*	*	*	*	*
Acetaminophen	*	*	*	*	*
Simvastatin Zocor	*	*	*	*	*
Atorvastatin Lipitor	*	*	*	*	*
Fluvastatin Lescol	*	*	*	*	*
Lovastatin Mevacor	*	*	*	*	*
Pravastatin Pravachol	*	*	*	*	*
Rosuvastatin Crestor	*	*	*	*	*
Cervastatin Baycol	*	*	*	*	*
Sulfonamide Antibiotics	*	*	*	*	*
Macrolide Antibiotics	*	*	*	*	*
Anti Fungal Azoles	*	*	*	*	*
Herbal Medications	*	*	*	*	*

Table 8: Contingency Analysis for the Koreans

Variable	Chinese vs. Japanese	Chinese vs. Malay	Chinese vs. Mixed	Chinese vs. NA
Gender	*		*	
Age	*	*	*	*
Ind. for Warfarin Treatment	*	*	*	*
Amiodarone Cordarone	*	*	*	*
Carbamazepine Tegretol	*	*	*	*
Phenytoin Dilantin	*	*	*	*
Rifampin or Rifampicin	*	*	*	*
CYP2C9 *1/*2			*	*
CYP2C9 *2/*2			*	
CYP2C9 *2/*3				*
CYP2C9_NA	*		*	*
VKORC1.1639_NA	*	*	*	*
VKORC1.1639_A/A	*	*	*	*
VKORC1.1639_A/G		*	*	*
VKORC1.497_NA	*	*	*	*
VKORC1.1173_C/C	*	*		*
VKORC1.1173_C/T	*	*	*	
VKORC1.1173_NA	*	*	*	*
VKORC1.1542_C/C	*	*	*	*
VKORC1.1542_C/G	*	*		*
VKORC1.1542_NA	*		*	*
VKORC1.3730_A/A	*	*		
VKORC1.3730_A/G	*	*		*
VKORC1.3730_NA	*		*	*
VKORC1.2255_C/C	*	*		
VKORC1.2255_C/T	*	*	*	*
VKORC1.2255_NA	*		*	*
VKORC1.4451_NA	*			*
Diabetes	*	*	*	
Congestive Heart Failure	*	*	*	*
Valve Replacement	*	*	*	*
Aspirin	*	*	*	
Acetaminophen	*	*	*	*
Simvastatin Zocor	*	*	*	*
Atorvastatin Lipitor	*	*	*	*
Fluvastatin Lescol	*	*	*	*
Lovastatin Mevacor	*	*	*	*
Pravastatin Pravachol	*	*	*	*
Rosuvastatin Crestor	*	*	*	*
Cervastatin Baycol	*	*	*	*
Sulfonamide Antibiotics	*	*	*	*
Macrolide Antibiotics	*	*	*	*
Anti Fungal Azoles	*	*	*	*
Herbal Medications	*	*	*	*

Table 9: Contingency Analysis for the Chinese

Variable	Japanese vs. Malay	Japanese vs. Mixed	Japanese vs. NA
Gender	*	*	*
Age		*	*
Ind. for Warfarin Treatment		*	*
Amiodarone Cordarone		*	
Carbamazepine Tegretol		*	
Phenytoin Dilantin		*	
Enzyme Inducer Status		*	*
CYP2C9 *1/*2			*
CYP2C9 *3/*3	*		
CYP2C9_NA	*		*
VKORC1_1639_NA	*	*	*
VKORC1_1639_A/A	*	*	*
VKORC1_1639_A/G		*	
VKORC1_497_G/G	*	*	*
VKORC1_497_NA	*	*	*
VKORC1_1173_C/C	*	*	*
VKORC1_1173_C/T		*	*
VKORC1_1173_NA	*	*	*
VKORC1_1542_C/C	*		*
VKORC1_1542_C/G		*	*
VKORC1_1542_NA	*	*	
VKORC1_3730_A/A	*		*
VKORC1_3730_A/G		*	*
VKORC1_3730_NA	*	*	*
VKORC1_2255_C/C	*		*
VKORC1_2255_C/T		*	*
VKORC1_4451_A/A			*
VKORC1_4451_A/C	*	*	*
VKORC1_4451_NA		*	*
Diabetes	*	*	*
Congestive Heart Failure	*	*	*
Valve Replacement		*	*
Aspirin		*	*
Acetaminophen		*	*
Simvastatin Zocor		*	
Atorvastatin Lipitor		*	
Fluvastatin Lescol		*	
Lovastatin Mevacor		*	
Pravastatin Pravachol		*	
Rosuvastatin Crestor		*	
Cervastatin Baycol		*	
Sulfonamide Antibiotics		*	
Macrolide Antibiotics		*	*
Anti Fungal Azoles		*	*
Herbal Medications	*	*	*

Table 10: Contingency Analysis for the Japanese

Variable	Malay vs. Mixed	Malay vs. NA	Mixed vs. NA
<b>Gender</b>			*
<b>Age</b>		*	*
<b>Ind. for Warfarin Treatment</b>	*	*	*
<b>Amiodarone Cordarone</b>	*	*	*
<b>Carbamazepine Tegretol</b>	*		*
<b>Phenytoin Dilantin</b>	*		*
<b>Rifampin or Rifampicin</b>	*		*
<b>CYP2C9 *1/*2</b>	*	*	
<b>CYP2C9_NA</b>	*	*	
<b>VKORC1.1639_NA</b>	*		*
<b>VKORC1.1639_A/A</b>	*		*
<b>VKORC1.1639_A/G</b>	*		*
<b>VKORC1.497_NA</b>	*		*
<b>VKORC1.1173_C/C</b>		*	*
<b>VKORC1.1173_C/T</b>	*	*	*
<b>VKORC1.1173_NA</b>	*	*	*
<b>VKORC1.1542_C/C</b>	*	*	
<b>VKORC1.1542_C/G</b>	*	*	*
<b>VKORC1.1542_NA</b>	*	*	*
<b>VKORC1.3730_A/A</b>		*	*
<b>VKORC1.3730_A/G</b>	*	*	*
<b>VKORC1.3730_NA</b>	*	*	*
<b>VKORC1.2255_C/C</b>		*	
<b>VKORC1.2255_C/T</b>	*	*	
<b>VKORC1.2255_NA</b>	*	*	
<b>Diabetes</b>	*	*	*
<b>Congestive Heart Failure</b>		*	*
<b>Valve Replacement</b>	*	*	*
<b>Aspirin</b>	*	*	*
<b>Acetaminophen</b>	*	*	*
<b>Simvastatin Zocor</b>	*	*	*
<b>Atorvastatin Lipitor</b>	*		*
<b>Fluvastatin Lescol</b>	*		*
<b>Lovastatin Mevacor</b>	*		*
<b>Pravastatin Pravachol</b>	*		*
<b>Rosuvastatin Crestor</b>	*		*
<b>Cervastatin Baycol</b>	*		*
<b>Sulfonamide Antibiotics</b>	*		*
<b>Macrolide Antibiotics</b>	*		*
<b>Anti Fungal Azoles</b>	*		*
<b>Herbal Medications</b>	*	*	*

Table 11: Contingency Analysis for the Malay and Mixed

### 3 Measure of performance of the NEJM algorithm by subgroups

Based on the many demographic differences between the subpopulations (as shown in Tables 7 - 11, we decided to evaluate the performance of the dosing algorithm previously described (reference the NEJM paper). The mean absolute error was used as the measure of fit, based on the published algorithm, shown in Figure 1 .

<b>Warfarin pharmacogenetic dosing algorithm</b>		
	5.6044	
-	0.2614 x	Age in decades
+	0.0087 x	Height in cm
+	0.0128 x	Weight in kg
-	0.8677 x	VKORC1 <sup>A</sup> /G
-	1.6974 x	VKORC1 A/A
-	0.4854 x	VKORC1 genotype unknown
-	0.5211 x	CYP2C9 *1/*2
-	0.9357 x	CYP2C9 *1/*3
-	1.0616 x	CYP2C9 *2/*2
-	1.9206 x	CYP2C9 *2/*3
-	2.3312 x	CYP2C9 *3/*3
-	0.2188 x	CYP2C9 genotype unknown
-	0.1092 x	Asian race
-	0.2760 x	Black or African American
-	0.1032 x	Missing or Mixed race
+	1.1816 x	Enzyme inducer status
-	0.5503 x	Amiodarone status
=	<b>Square root of weekly warfarin dose**</b>	

Figure 1: published NEJM algorithm

The Mean Absolute Error (MAE) is a measure of the performance of a forecasting model and is expressed as:

$$\begin{aligned} MAE &= \frac{1}{n} \sum_{i=1}^n |f_i - y_i| \\ &= \frac{1}{n} \sum_{i=1}^n |e_i| \end{aligned}$$

Where

- $n$  is the size of the dataset used.
- $i$  is the index row indicator of individuals.
- $f_i$  is the vector of forecasted values by the algorithm.
- $y_i$  is the vector of observed outcomes in the dataset.
- $e_i$  is the vector of errors of the model.

For each stratified group, the algorithm was used to generate predicted values for the square root transformed therapeutic dose of warfarin. These predicted values were used in calculating the MAE and standard deviation of these errors for each group (shown in Table 1).

Group	N Rows	MAE (Mean/Std Dev)
<b>Total</b>	5103	0.840 (0.742)
<b>OMB_White</b>	3035	0.840 (0.764)
<b>OMB_Black</b>	663	0.979 (0.861)
<b>OMB_Asian</b>	958	0.786 (0.627)
<b>OMB_Other</b>	13	0.666 (0.454)
<b>Korean</b>	259	0.727 (0.486)
<b>Chinese</b>	335	0.761 (0.601)
<b>Japanese</b>	227	0.786 (0.707)
<b>Malay</b>	82	0.983 (0.735)
<b>Mixed</b>	80	0.679 (0.532)
<b>NA</b>	464	0.834 (0.612)

Table 12: NEJM algorithm [6] Performance

This chart shows the MAE for each subpopulation. In addition, this table shows the standard deviation in parenthesis. The main observation of the table is that the MAE observed is similar between OMB groups with more accuracy observed in the OMB Asian group. The variability observed is also very uniform between groups with only one exception in the Korean case which is less variable (0.486).

## 4 Dosing Predictions for the Anticoagulant Warfarin

### 4.1 Univariate analysis

Univariate least square regression analyses were conducted for the whole data set and each subpopulation to determine the effect of each variable separately. Since the distribution of the therapeutic dose of warfarin was right skewed, the square-root transformation of doses was used. One patient with a weekly dose of 315 mg was excluded in the analyses because this dose was considered extraordinarily high and rare. Table 13

and 14 shows the univariate analysis results, where  $\star$  indicates a significant association between the predictor and the outcome at  $\alpha$  level of 0.05. The tables also shows how the predictors affect the outcome differently across subpopulations. For example, patient's *height* is significantly associated with the outcome for the whole population and all the subpopulations except for the Chinese and Malay. The patient's *weight* is significantly associated with the outcome for the whole population and all the subpopulations except for the Korean. The presence of genotype variants of CYP2C9 affects the outcome for the whole population, the OMB white, OMB Asian, and OMB Other group, but not for the OMB Black group. It also affects the outcome for the Chinese, Japanese, and NA group, but not for the Korean, Malay, and Mixed group. All seven single nucleotide polymorphisms (SNPs) of the gene VKORC1 have significant impact on warfarin dose for the whole population and the OMB White subpopulation. For other subpopulations, at least one of the seven SNPs are associated with the outcome. It is also worth mentioning that the VKORC1 -4451 consensus SNP has impact on only the OMB White subpopulation.

Variable	Whole	OMB White	OMB Black	OMB Asian	OMB Other
Height (cm)	*	*	*	*	*
Weight (kg)	*	*	*	*	*
Target INR	*				*
INR on Reported Therapeutic Dose of Warfarin	*				
Age	*	*	*	*	*
Gender	*	*			*
Indication for Warfarin Treatment	*	*	*	*	
Amiodarone (Cordarone)	*	*	*	*	*
Carbamazepine (Tegretol)	*	*			
Phenytoin (Dilantin)	*	*			
Rifampin or Rifampicin	*	*			
Enzyme inducer status	*	*			
CYP2C9	*	*		*	*
VKORC1 -1639 consensus	*	*	*	*	*
VKORC1 497 consensus	*	*	*	*	
VKORC1 1173 consensus	*	*	*	*	*
VKORC1 1542 consensus	*	*	*	*	*
VKORC1 3730 consensus	*	*	*	*	*
VKORC1 2255 consensus	*	*		*	*
VKORC1 -4451 consensus	*	*			
Diabetes	*				
Congestive Heart Failure and/or Cardiomyopathy	*		*	*	
Valve Replacement	*	*		*	*
Aspirin			*	*	
Acetaminophen or Paracetamol (Tylenol)	*				
Herbal Medications, Vitamins, Supplements	*	*		*	

Table 13: Univariate Analysis for OMB Races

Variable	Korean	Chinese	Japanese	Malay	Mixed	NA
Height (cm)		*	*		*	*
Weight (kg)	*	*		*	*	*
Target INR					*	
INR on Reported Therapeutic Dose of Warfarin		*				*
Age	*	*		*		*
Gender				*		*
Indication for Warfarin Treatment					*	*
Amiodarone (Cordarone)						
Carbamazepine (Tegretol)						
Phenytoin (Dilantin)						
Rifampin or Rifampicin						
Enzyme inducer status						*
CYP2C9	*	*				*
VKORC1 -1639 consensus	*	*	*	*	*	
VKORC1 497 consensus		*				
VKORC1 1173 consensus	*	*	*	*		*
VKORC1 1542 consensus		*	*	*		*
VKORC1 3730 consensus	*	*	*	*		
VKORC1 2255 consensus		*	*	*		
VKORC1 -4451 consensus						
Diabetes						
Congestive Heart Failure and/or Cardiomyopathy						
Valve Replacement					*	*
Aspirin	*					
Acetaminophen or Paracetamol (Tylenol)						
Herbal Medications, Vitamins, Supplements						

Table 14: Univariate Analysis for the Asian Subgroups

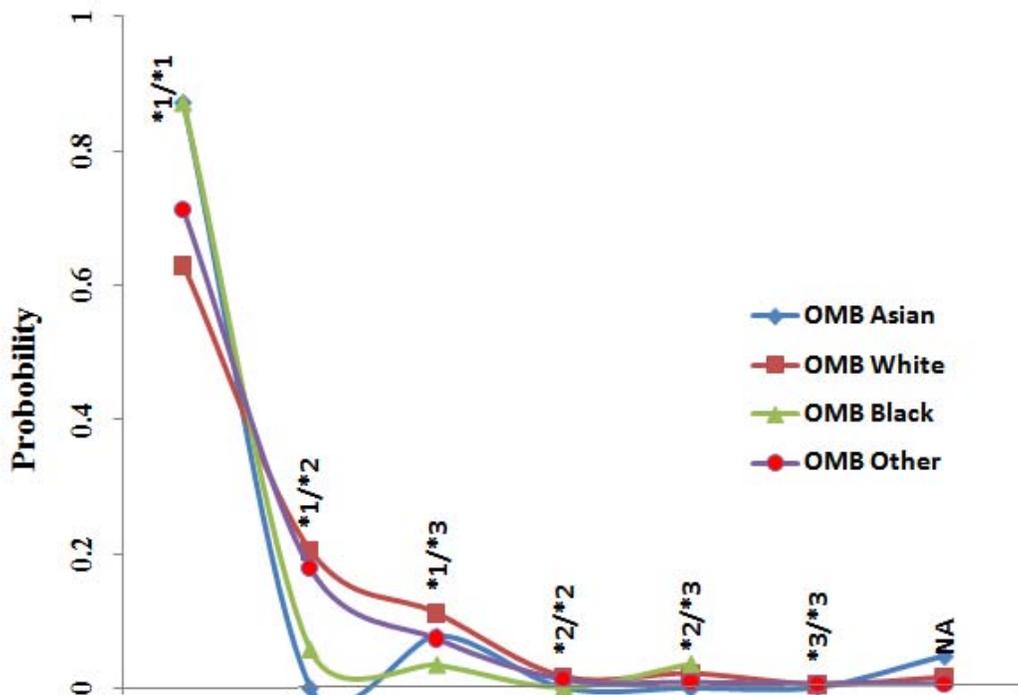


Figure 2: Frequency of different CYP2C9 variants in OMB population groups

As the gene effects on the Warfarin dosing is the focus of this study, a precise knowledge about the distribution of different gene variants within the various races is required. In Figure 2 the frequency of different CYP2C9 variants (represented as proportions/probabilities), namely \*1/\*1, \*1/\*2, \*1/\*3, \*2/\*2, \*2/\*3, \*3/\*3 among the racial groups is shown. It can be seen that the OMB White and Black groups are significantly different in the probability of CYP2C9 \*1/\*1. In addition, the OMB Asian and White show different behavior for the CYP2C9 \*1/\*2 distribution.

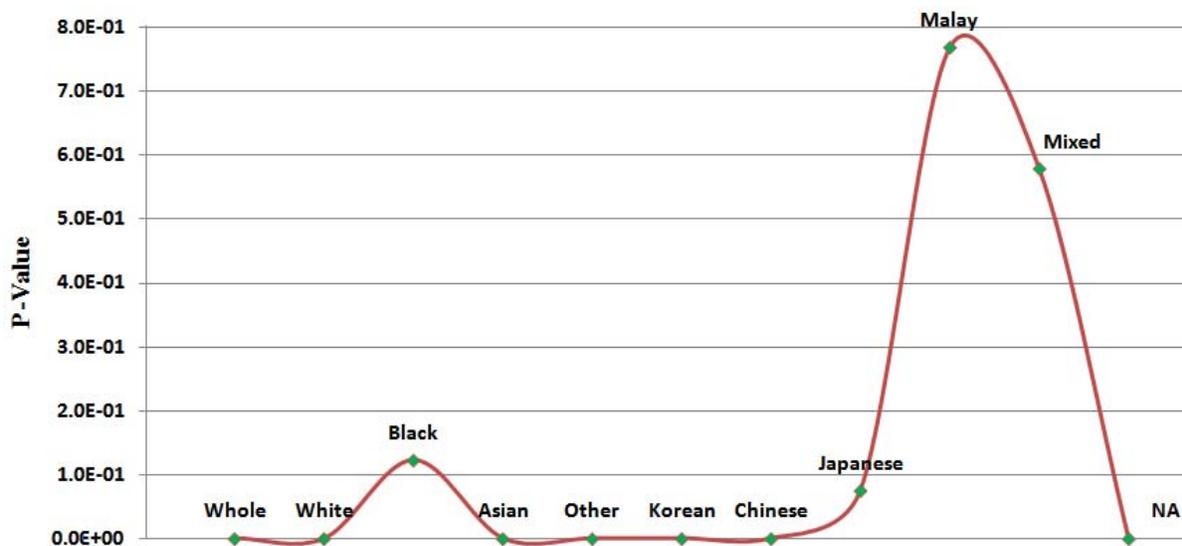


Figure 3: Univariate analysis of CYP2C9 for different subpopulations

Figure 3 shows the results of the univariate analysis of CYP2C9 for different subpopulations. As seen in this Figure, CYP2C9 does not have a strong effect on the outcome for some subpopulations. It also can be seen that the idea of splitting the OMB subpopulations, especially the OMB Asian group, to smaller race/ethnic groups allows us to see the difference of the effect of the genetic factors across these refined sub-groups. For example, the effect of CYP2C9 is substantially different between the Japanese and the Malay group, which were included in the OMB Asian group.

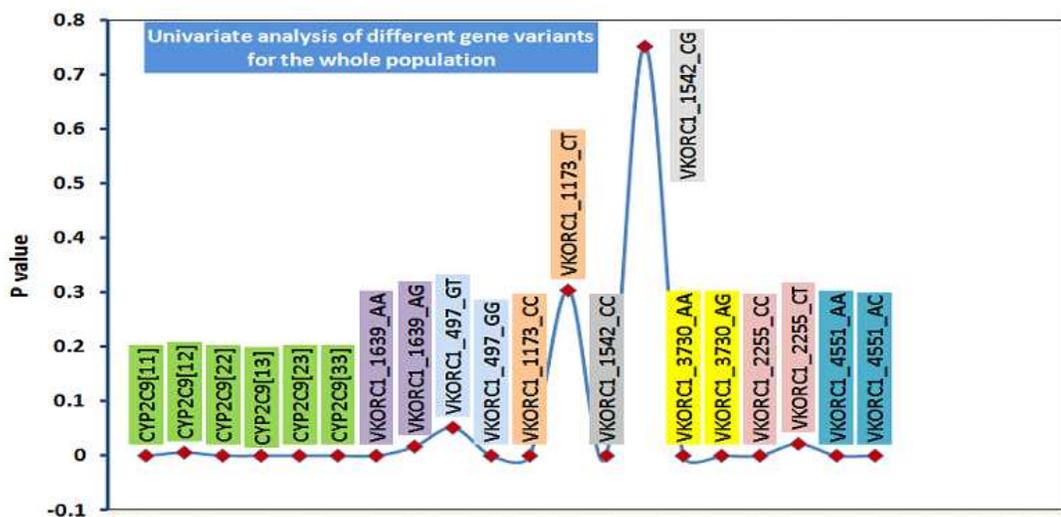


Figure 4: Univariate analysis of different gene variants for the whole population

Figure 4 shows the p-values from the univariate analysis for the whole population with dummy coded variants of CYP2C9 and VKORC1 SNPs. It can be seen that VKORC1 497 GT, VKORC1 113 CT and VKORC1 1542 CG do not have an important effect on dosing prediction.

## 4.2 Whole Population And Subpopulation Models

Multivariable least square regression models were run for the whole population and each of the subpopulations. Again, square-root transformed doses were used, and the single large outlier was excluded in the analyses. Variable selection was carried out with the backward selection method. The significance level for staying in the model was chosen as 0.05. Observations with missing on *age*, *height*, *weight*, or *gender* were excluded in the analyses. For all other variables, missing was treated as a distinct category. Dummy variables were created for each of the class variables with more than two categories. Table 15 shows the final models and parameter estimates. It can be seen that the subpopulation models are quite different from each other and from the model based on the whole population. After adjusting for other variables in the model, the *age*, *height*, and *weight* factors act more or less the same for the OMB groups, but not for the more detailed race/ethnic groups. Other demographic and background factors act differently across subpopulations. The variations in the two genes, CYP2C9 and VKORC1, have similar effect for the whole population and the OMB White subpopulation but not for other subpopulations. This may due to the fact that more than half of the sample in our study is Caucasian, and for both genes, the haplotype and allele frequencies are different among stratified subpopulations.

To evaluate the predictive accuracy of each model, the absolute difference between the predicted and observed dose (in the square-root scale), i.e. absolute error, was calculated for each patient, and the mean absolute error (MAE) was obtained for each model. Table 16 shows the MAE and its standard deviation, the mean predicted weekly dose in the square-root scale and its standard deviation, and the coefficient of determination  $R^2$ , which is defined here as the square of the sample correlation coefficient of the predicted dose and

the observed dose (in the square-root scale). The results show that the MAEs from the subpopulation models are consistently lower than the ones from the whole population model, and the  $R^2$ 's from the subpopulation models are consistently higher than the ones from the whole population model. Figure 5 shows the box plots of the observed dose minus the predicted dose, i.e. residuals. On average, compared to the whole population model, the subpopulation models give residuals with less extreme values and a narrower interquartile range. We were unable to build a decent model for the Japanese subpopulation because most of the background variables were missing.

Variable	OMB										
	Whole	White	Black	Asian	Other	Chinese	Japanese	Korean	Malay	Mixed	NA
Intercept	4.38	4.37	3.21	7.22	2.17	7.03	4.91	5.21	7.82	5.72	2.30
Age in decades	-0.22	-0.25	-0.20	-0.22	-0.20	-0.26		-0.13	-0.35	-0.18	-0.26
Height (cm)	0.01	0.02	0.02	-0.01	0.02						0.02
Weight (kg)	0.01	0.01	0.01	0.02	0.01	0.01		0.01		0.02	0.01
Indication 1	0.27	0.23									1.79
Indication 2	0.25	0.20									1.63
Indication 3										-0.62	1.38
Indication 4	0.22	0.43				-0.31					2.06
Indication 7	0.35	0.37									
Indication 8	0.16			0.26		0.51					1.71
Acetaminophen NA	-0.27	-0.29									
Amiodarone	-0.56	-0.56	-0.85	-0.50	-0.81	-0.65		-0.66			-1.07
Amiodarone NA	-0.18										
Aspirin								-0.55			
Atorvastatin	-0.22				1.22						
Atorvastatin NA	-0.67										
Carbamazepine								1.22			1.83
Cerivastatin NA	1.32	1.99	2.81								
Congestive Heart Failure	-0.18	-0.14	-0.28	-0.23	1.48						
Congestive Heart Failure NA	0.68	1.70	1.30		0.84						
Diabetes	0.10									1.18	
Diabetes NA	0.25										
Enzyme Inducer Status	1.18	1.17									
Fluvastatin			-2.75								
Fluvastatin NA	-0.42										
Herbal NA	0.13	0.16									
Lovastatin	-0.46										
Phenytoin	-0.54		1.03								
Phenytoin NA			-1.64								
Rosuvastatin NA		-1.80									
Simvastatin	-0.12										
Simvastatin NA	0.30										
Sulfonamide NA						0.39					
Valve Replacement						0.38					
Valve Replacement NA	-1.24	-1.96	-2.60								
CYP2C9 12	-0.48	-0.51			-0.34						-0.39
CYP2C9 13	-0.89	-0.99	-0.57	-0.85	-0.69	-1.02		-1.02			-0.86
CYP2C9 22	-1.13	-1.19			-0.95						-1.73
CYP2C9 23	-1.74	-1.76			-2.50						-2.35
CYP2C9 33	-2.45	-2.61		-2.62	-1.49						-3.60
CYP2C9 NA	-0.36	-0.65									
VKORC1 -1639 AA	-1.00	-1.08	-1.14		-1.78		-0.62			-1.44	
VKORC1 -1639 AG	-0.52	-0.50	-0.83		-0.72					-0.48	
VKORC1 -1639 NA	-0.29	-0.35	-0.69		-1.36						
VKORC1 497 NA	-0.23				1.36						
VKORC1 1173 CC	0.86	1.00		1.22	1.49						1.55
VKORC1 1173 CT	0.48	0.58		0.69	0.71			0.85			0.75
VKORC1 1173 NA	0.24	0.34	0.33								
VKORC1 1542 CC	-0.77	-0.96		-0.71		-2.14			-1.67		-1.59
VKORC1 1542 CG	-0.41	-0.47				-1.23					-0.69
VKORC1 1542 NA	-0.23	-0.37			-1.74	-1.69					
VKORC1 3730 AA											-1.43
VKORC1 3730 AG									-0.77		
VKORC1 3730 NA	-0.33	-0.34									
VKORC1 2255 CC	-0.37	-0.79									
VKORC1 2255 CT	-0.27	-0.52		-0.58							
VKORC1 2255 NA	0.30				1.48						
VKORC1 -4451 AA											-1.98
VKORC1 -4451 AC											-1.05
VKORC1 -4451 NA				-0.25							
OMB Other	-0.49										
OMB Black	-0.37										

Table 15: Multivariable Analyses

	Pred Sqrt Dose		MAE		$R^2$ (%)	
	Sub-Pop Model	Whole Pop Model	Sub-Pop Model	Whole Pop Model	Sub-Pop Model	Whole Pop Model
Whole	-	5.56 (0.98)	-	0.77 (0.66)	-	48
OMB White	5.62 (1.00)	5.63 (0.94)	0.77 (0.66)	0.78 (0.67)	49	48
OMB Black	6.28 (0.95)	6.28 (0.89)	0.89 (0.73)	0.95 (0.77)	41	33
OMB Asian	4.85 (0.73)	4.82 (0.77)	0.59 (0.52)	0.64 (0.56)	46	39
OMB Other	5.53 (0.86)	5.53 (0.82)	0.65 (0.53)	0.74 (0.59)	51	39
Chinese	4.66 (0.73)	4.64 (0.77)	0.56 (0.47)	0.62 (0.53)	50	40
Japanese	4.36 (0.20)	4.47 (0.54)	0.72 (0.62)	0.74 (0.69)	4	5
Korean	5.27 (0.51)	5.17 (0.62)	0.46 (0.38)	0.52 (0.43)	42	33
Malay	4.97 (0.71)	5.10 (0.66)	0.55 (0.39)	0.64 (0.47)	53	37
Mixed	5.60 (0.67)	5.77 (0.77)	0.58 (0.41)	0.69 (0.54)	46	29
NA	5.58 (0.91)	5.49 (0.75)	0.66 (0.55)	0.73 (0.58)	53	44

Table 16: Model performance

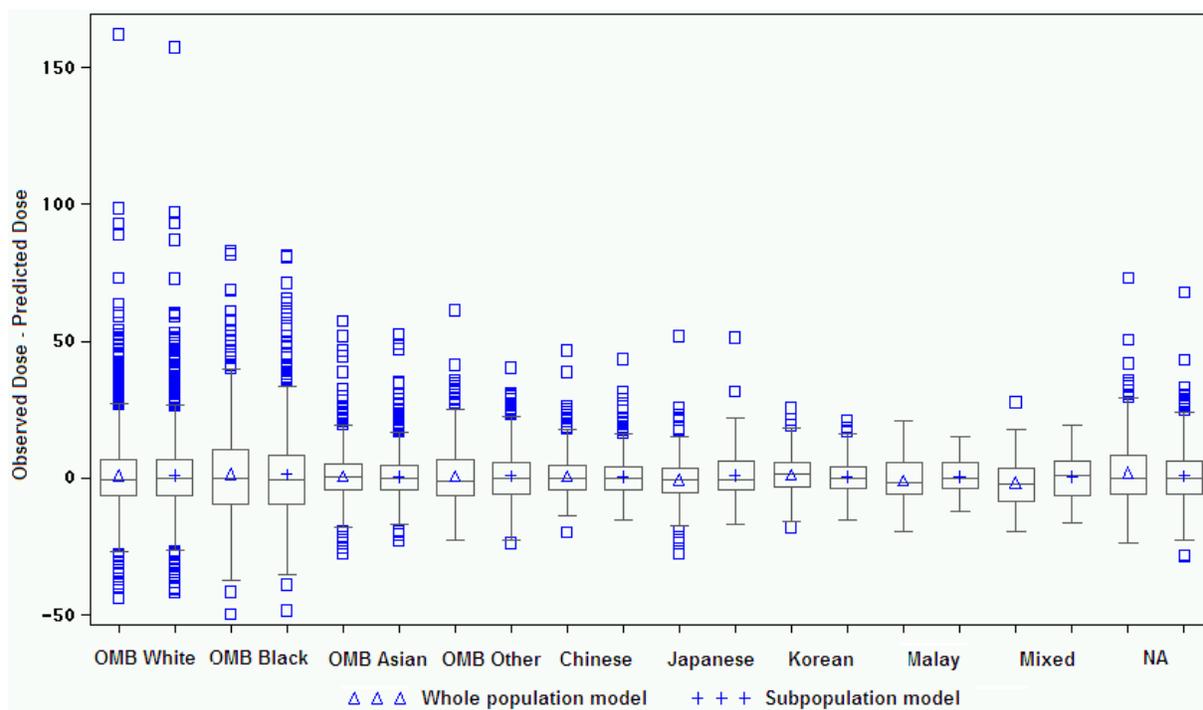


Figure 5: Model performance by race

## 5 Summary

In the current project, we demonstrated there are potentially important demographic characteristics (both clinical and genetic factors) in the racial and ethnic subgroups and evaluated the performance of the previously published dosing prediction in each of these subgroups. As expected, there were differences in the performance between the subgroups, with the smallest errors in the Asian groups and the largest error in the Black group.

Because of these differences in performance, we performed model building to develop of new dose prediction

algorithm specifically for each subpopulation. These models were built using linear regression with backwards selection for variable selection. The model fit and performance for each of these newly constructed models was calculated and compared to the models based on the entire population. Our results indicate improved model fit and predictive performance for these new models compared to a global model. While we compared measure of model fit and prediction statistically, the real value of these models should be evaluated in future studies. [1] Evaluating these models for clinical utility in future prospective and observational studies would provide the most important measures of model fit - clinical utility.

Additionally, while these models have extremely high predictive performance (with 40-50% of the variation in warfarin dosing explained), there is still a large amount of variation that is not explained by these models. As additional factors that influence warfarin dose are identified, these models should be re-evaluated and updated to obtain even better predictions to further increase the efficacy and decrease the risks associated with warfarin use.

## 6 Acknowledgments

We would like to thank the International Warfarin Pharmacogenetics Consortium (IWPC) for providing the data used here. We would also like to thank the patients that contributed to the IWPC data. In addition, we would like to thank the organizers and our advisors.

## References

- [1] Attia J, Ioannidis JP, Thakkinstian A, et al. How to use an article about genetic association. *JAMA*. 2009; 301(1):74-81.
- [2] Caldwell M, Awad T, Johnson J, Gage B et al. CYP4F2 genetic variant alters required warfarin dose. *Blood* 2008.
- [3] Caraco Y, Blotnick S, Muszkat M. CYP2C9 genotype-guided warfarin prescribing enhances the efficacy and safety of anticoagulation: a prospective randomized controlled study. *Clin Pharmacol Ther* 2008:460-70.
- [4] Gage B, Eby C, Johnson J, Deych E et al. Use of Pharmacogenetic and Clinical Factors to Predict the Therapeutic Dose of Warfarin. *Clin Pharmacol Ther* 2008.
- [5] Jonas DE, McLeod HL. Genetic and clinical factors relating to warfarin dosing. *Trends in Pharmacological Sciences*;30(7):375-86.
- [6] L. Lamport, LaTeX:A Document Preparation System, Addison-Wesley Publishing Company, New York, 1994.
- [7] The International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine* 2009;360:753-64.

# SEVERE WEATHER UNDER A CHANGING CLIMATE: LARGE-SCALE INDICATORS OF EXTREME EVENTS

Matthew Heaton<sup>1</sup>, Matthias Katzfuß<sup>2</sup>, Yi Li<sup>3</sup>, Kathryn Pedings<sup>4</sup>, Shahla Ramachandar<sup>5</sup>

## **Problem Presenter:**

Dr. Eric Gilleland  
National Center for Atmospheric Research

## **Faculty Mentors:**

Dr. Elizabeth Mannshardt-Shamseldin - SAMSI / Duke University  
Dr. Richard L. Smith - University of North Carolina at Chapel Hill

July 28, 2009

## **Abstract**

Global warming has become a generally accepted theory in the past decade or so. A recent wave of severe storms has raised the question of whether global warming has or will result in an increase in both frequency and intensity of such extreme weather events. This paper will examine large-scale indicators that are generally believed to be closely related to severe storms, using extreme value theory. We present numerous methods, such as the Generalized Extreme Value distribution, the Generalized Pareto distribution, and Bayesian approaches incorporating spatial correlation between measurement locations. These approaches are applied to a data set covering the United States and surrounding regions for the years 1958 to 1999, to predict return levels for extreme events. We also attempt to determine if there was a change in the behavior of these events over the time period of interest.

---

<sup>1</sup>Duke University, Department of Statistical Sciences, matt@stat.duke.edu

<sup>2</sup>The Ohio State University, Department of Statistics, katzfuss@stat.osu.edu

<sup>3</sup>Duke University, Department of Mathematics, yi.li@duke.edu

<sup>4</sup>College of Charleston, Department of Mathematics, kepeding@edisto.cofc.edu

<sup>5</sup>University of Texas - Dallas, Department of Mathematical Sciences, shahla@utdallas.edu

# Contents

<b>1</b>	<b>Introduction and Motivation</b>	<b>3</b>
<b>2</b>	<b>Extreme Value Theory</b>	<b>4</b>
2.1	GEV distribution . . . . .	4
2.2	GPD . . . . .	5
<b>3</b>	<b>Exploratory Data Analysis</b>	<b>5</b>
3.1	Data . . . . .	5
3.2	Stationary Models . . . . .	7
3.2.1	GEV . . . . .	7
3.2.2	GPD . . . . .	8
3.2.3	Comparison of GEV and GPD . . . . .	9
3.3	Declustering . . . . .	11
3.4	Time Trend . . . . .	12
<b>4</b>	<b>Analysis using Bayesian Spatial Extreme Value Models</b>	<b>13</b>
4.1	Model 1 . . . . .	14
4.2	Model 2 . . . . .	15
4.3	Model Comparison . . . . .	16
4.4	Analysis of Temporal Trend . . . . .	16
<b>5</b>	<b>Conclusion and Future Work</b>	<b>17</b>

# 1 Introduction and Motivation

Changing weather patterns have always been of interest, but what is of greater concern is the behavior of extreme events. These pose significant threats to life and property. In many cases, they have a large economic impact as well. An unexpected high-intensity tornado or snowfall could have serious repercussions. For instance, the 20.3-inch snowfall in Raleigh, NC of 2000 [18] or the infamous hurricanes in the recent past. In this paper, we look specifically at severe storms and whether they are becoming more severe or frequent over time. Storms can be classified by windspeed (in kts.) and size of hail (diameter in cms.)

Severity	Classifiers
Non-severe	Hail < 1.9 and winds < 55 , no tornado
Severe	Hail ≥ 1.9 and 55 ≤ wind < 65 or tornado
Significant non-tornadic	Hail ≥ 5.07 and wind ≥ 65
Significant tornadic	Same as above but with a tornado of intensity ≥ F2

Table 1: Table of storm classification [14].

Presently the average time of advance warning for an approaching tornado is 5-13 minutes [12]. The coarse resolution of current climate models renders them incapable of capturing fine scale extreme events such as tornadoes or hurricanes. This leads us into the consideration of large-scale indicators which can predict extreme weather much farther in advance. Of specific interest are the convective available potential energy (CAPE), measured in Joules/Kg, and the vertical shear of the tropospheric horizontal winds (WS), measured in meters/second. From [1], a combination of these two parameters results in a good indicator of extreme weather events. Typically, severe thunderstorms are associated with large values of CAPE and WS.

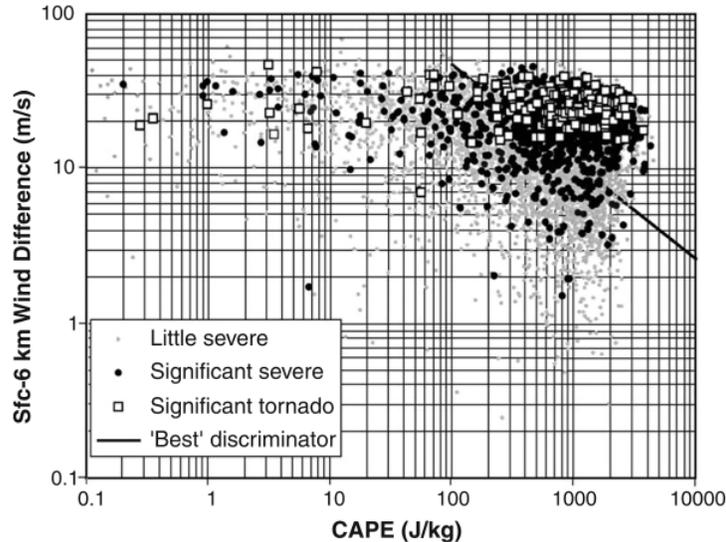


Figure 1: A scatterplot of CAPE and WS associated with severe thunderstorms in the US from 1997 - 1999 [1].

We consider

$$W_{max} = \sqrt{(2 \times CAPE)}$$

$$WmSh = W_{max} \times WS$$

Scaling  $W_{max}$  in this way puts the indicator on the same scale as windshear, in meters per second.

The  $n$ -year return level,  $y_n$ , is the extreme level expected to occur every  $n$  years. This paper deals with the 20-year return levels of  $WmSh$ . Note that the return levels are not for storms themselves, but for an indicator that tells whether the environment is conducive for severe storms or not.

Our main goal is to try to provide probabilistic information about storms' frequency and intensity of occurring. This study begins with a brief introduction to univariate extreme value theory. We discuss the *Generalized Extreme Value* (GEV) distribution and the *Generalized Pareto Distribution* (GPD) in sections 2.1 and 2.2. The data are described in section 3.1. In section 3.2, an exploratory analysis is done using the GEV and the GPD methods, immediately followed by a comparison of the two methods based on results from our data set.

Next, we briefly address the issue of dependence between subsequent extreme values in section 3.3. High threshold exceedances tend to occur in clusters over time. We look into whether declustering may provide some reasonable solutions to this problem. As a possible improvement to the GEV method, in section 3.4 we attempt to incorporate a time component, or linear trend into the GEV model.

Further, in section 4 we consider employing Bayesian methods and spatial aspects in the analysis. Under Bayesian methods, a prior density function for the unknown parameters is specified, and then a posterior density for the parameters, given the observations, is computed. We account for the fact that extremal behavior may also be related to a spatial component, included through the covariates latitude and longitude. The Markov chain Monte Carlo (MCMC) methods are used for these computations.

## 2 Extreme Value Theory

This paper studies methods for univariate extreme value analysis. This section describes the two basic methods, namely, the GEV distribution and the GPD, which we have employed to do the analysis. More details can be found in [18].

### 2.1 GEV distribution

Let  $X_1, X_2, \dots, X_n$  be a sequence of i.i.d. random variables with distribution function  $F$ . Suppose  $M_n = \max(X_1, \dots, X_n)$  is the sample maximum. Then we know that

$$Pr \{M_n \leq x\} = F(x)^n,$$

so for any fixed  $x$  such that  $F(x) < 1$ ,  $Pr \{M_n \leq x\} \xrightarrow{n \rightarrow \infty} 0$

To obtain a non-trivial limit result, we must renormalize. We need to find sequences  $a_n$  and  $b_n$  such that

$$Pr \left\{ \frac{M_n - b_n}{a_n} \leq x \right\} = F(a_n x + b_n)^n \rightarrow H(x)$$

Here we invoke the *Three Types Theorem* first stated by Fisher and Tippett (1928) and later proved by Gnedenko (1943):

If a non-degenerate  $H$  exists, then it must be one of the following three types:

$$H(x) = \exp(-e^{-x}), \text{ for all } x \tag{1}$$

$$H(x) = \begin{cases} 0, & x < 0, \\ \exp(-x^{-\alpha}), & x > 0 \end{cases} \tag{2}$$

$$H(x) = \begin{cases} \exp(-|x|^{-\alpha}), & x < 0 \\ 1, & x > 0 \end{cases} \tag{3}$$

Equation (1) is called the *Gumbel* type, equation (2) the *Fréchet* type and equation (3) the *Weibull* type. The three distributions may be combined into a single type called the GEV distribution. The GEV distribution is defined by

$$P(X \leq x) = \exp \left\{ - \left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\},$$

where  $x_+ = \max(x, 0)$ . In this equation,  $\mu$  is the location parameter,  $\sigma$  is the scale parameter and  $\xi$  is the shape parameter. It can be seen immediately that the limit  $\xi \rightarrow 0$  corresponds to the Gumbel type,  $\xi > 0$  to the Fréchet type and  $\xi < 0$  to the Weibull type.

## 2.2 GPD

Long-tailed processes have been modeled using the Pareto distribution for some time now. A rigorous connection of the GPD with classical extreme value theory was established by Pickands (1975).

Let  $X_1, X_2, \dots, X_n$  be a sequence of i.i.d. random variables with distribution function  $F$ . In this method, first a threshold  $u$  is fixed. The threshold should be high enough to ensure that only the truly extreme values are included in the analysis but not so high so as to prevent proper model-fitting. Threshold selection can be a topic for further exploration but is not studied fully here. All points  $X_i$  in the data above this threshold are included in the analysis. Consider the conditional distribution of the exceedances  $Y_i = X_i - u$ :

$$Pr \{Y \leq y | Y > 0\} = \frac{F(u + y) - F(u)}{1 - F(u)}$$

As  $u \rightarrow \sup \{x : F(x) < 1\}$ , we have

$$Pr \{Y \leq y | Y > 0\} \rightarrow G(y, \tilde{\sigma}, \xi),$$

where  $G$  is the GPD given by

$$G(y, \tilde{\sigma}, \xi) = 1 - \left(1 + \xi \frac{y}{\tilde{\sigma}}\right)_+^{-1/\xi}.$$

Here  $\tilde{\sigma}$  and  $\xi$  are the scale and shape parameter, respectively. We can establish a relation between these parameters and the parameters of the GEV distribution as follows:

$$\xi_{GPD} = \xi_{GEV} = \xi$$

$$\tilde{\sigma} = \sigma + \xi(u - \mu)$$

The main advantage of the GPD method is that more observations can be used than for the GEV distribution, which only considers the annual maxima. This usually leads to more precise estimates of parameters and other statistics, e.g., return levels.

## 3 Exploratory Data Analysis

### 3.1 Data

The data used is part of Brooks' global data set [1]. The re-analysis data has all observations from each time index synthesized using a static data assimilator. The *WmSh* data set consists of daily observations at 884 different locations on a  $1.4 \times 1.4$  degree grid across the United States and adjacent regions over a 42-year time period (1958-1999). As several days are not observed, we have 15,300 observations at each location. For a description of re-analysis data, see [1]. The data set used for this report has only one variable, namely, the *WmSh* value itself. For more information, refer to [15].

Figure 2 is a simple histogram of *WmSh* values for  $1000 < WmSh < 5500$ .

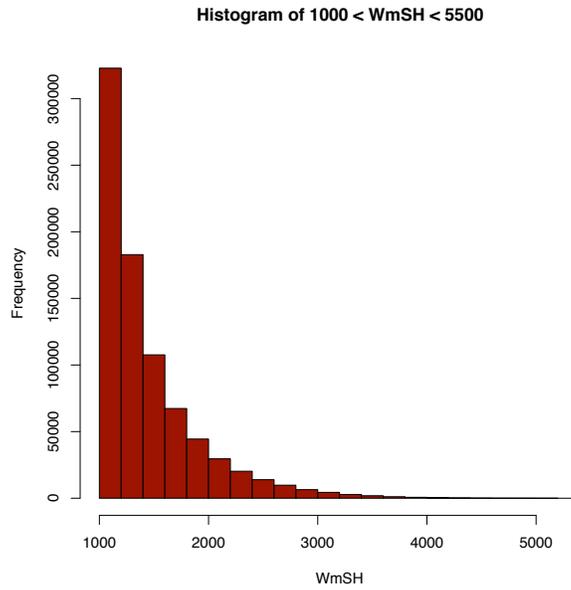


Figure 2: Histogram of  $WmSh$  Data for 1958 - 1999.

Extreme value theory is not interested in all of the data, only the extreme points. Figure 3 shows the spatial distribution of the data, in form of the three quartiles and the 95<sup>th</sup> percentile, which will be used later (section 3.2.2). It is important to note that the scale for the 95<sup>th</sup> percentile is significantly higher than those of the other quartiles.

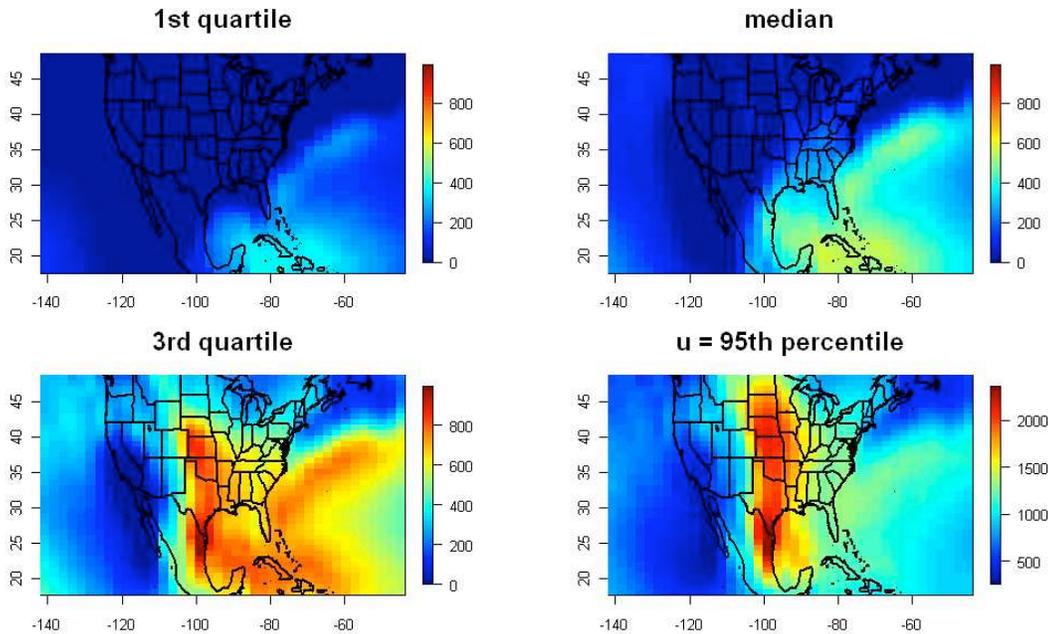


Figure 3: Heat Maps of each of the Quartiles and (on a different scale) the 95<sup>th</sup> Percentile for the Data

## 3.2 Stationary Models

### 3.2.1 GEV

In order to use the GEV method for our data, we first extract the annual maximum at each location and then we obtain the maximum likelihood estimates of the three parameters of the GEV distribution,  $\mu$ ,  $\sigma$  and  $\xi$ , for each of the 884 locations. Figure 4 shows the results we obtained using functions in R [2, 3, 7, 8, 11, 20, 10].

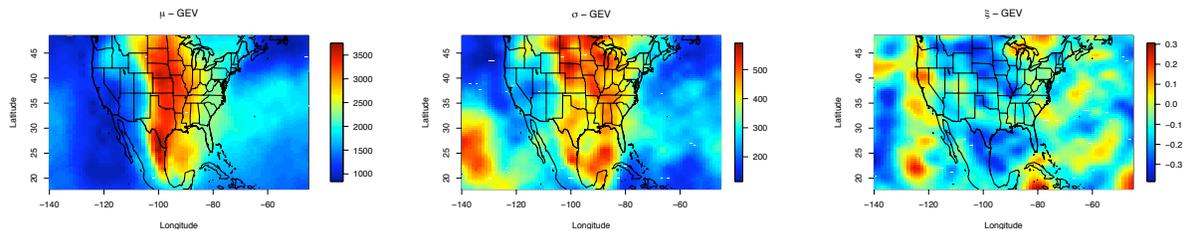


Figure 4: Heat Plots of  $\mu$ ,  $\sigma$ , and  $\xi$  Values of the GEV Distribution

The location parameter  $\mu$  anchors the distribution towards the mean of extreme values of  $WmSh$ . Estimates of  $\mu$  give us an idea of the average of extreme  $WmSh$  values across the region under consideration. In the first plot, we can see a concentration of high values along the ‘tornado alley’ of the United States and downward towards parts of Mexico and the Gulf of Mexico. (For the purpose of this paper, we will refer to this entire region as ‘tornado alley’.)

The scale parameter  $\sigma$  tells us about the spread of the data, but is not the same as the standard deviation. From the plot we see that the data seem to be more variable in tornado alley. We also observe this phenomenon over a part of the Pacific Ocean in the bottom left of the graph. This may be a consequence of the fact that data collection over the ocean might not be as reliable as that over land. However, due to time constraints, further analysis of this issue was not done.

Estimated values of the shape parameter  $\xi$  along the tornado alley seem to be negative, ranging from just below 0 to about -0.4. A negative  $\xi$  corresponds to the Weibull distribution from section 2.1. This could be due to the high values of  $WmSh$  in this area, so that many observations could have come close to the physical boundary of the underlying process. Although they are negative, all values of  $\xi$  are within acceptable reasonable range for MLE estimation. The remaining areas, however, show predominantly positive values, ranging from just above 0 to about 0.3.

After obtaining the maximum likelihood estimates, we can obtain a plug-in estimator of the 20-year return levels. Figure 5 shows the 20-year return level for each spatial location. Once again, values are particularly high in tornado alley, which is in accordance with what we would expect.

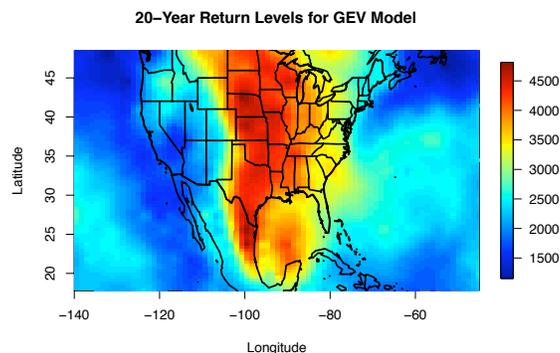


Figure 5: Heat Plot of the 20-Year Return Levels of the GEV Distribution

To assess the goodness of fit, we look at several diagnostic plots. Figure 6 shows quantile-quantile (Q-Q) plots and histograms for two randomly chosen locations, indicating a reasonable fit of the GEV distribution to the data. The first site is over the Pacific, while the second one is located in the northeast United States.

Note that the first site does not contain many extreme values, and so the GEV distribution can easily account for this. Due to the many extreme values at the second site, the data fit is not as good.

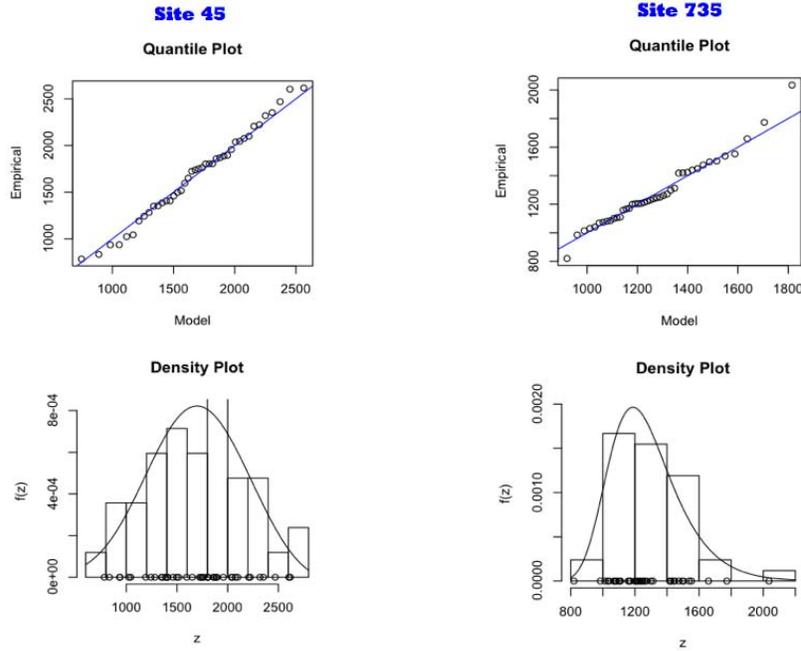


Figure 6: Q-Q and Histograms of sites 45 and 735 for GEV

### 3.2.2 GPD

As stated in section 2.2, to obtain the exceedances that go into the GPD distribution, we first need to select a threshold  $u$ . Having a fixed  $u$  across all locations would be problematic, as the location of the data distribution is very different across the region of interest (see Figure 3). Therefore, we choose  $u$  to be the 95<sup>th</sup> percentile at each location. That is, at each station separately, we include the top 5% of values. The bottom right plot in Figure 3 shows these 95<sup>th</sup> percentiles.

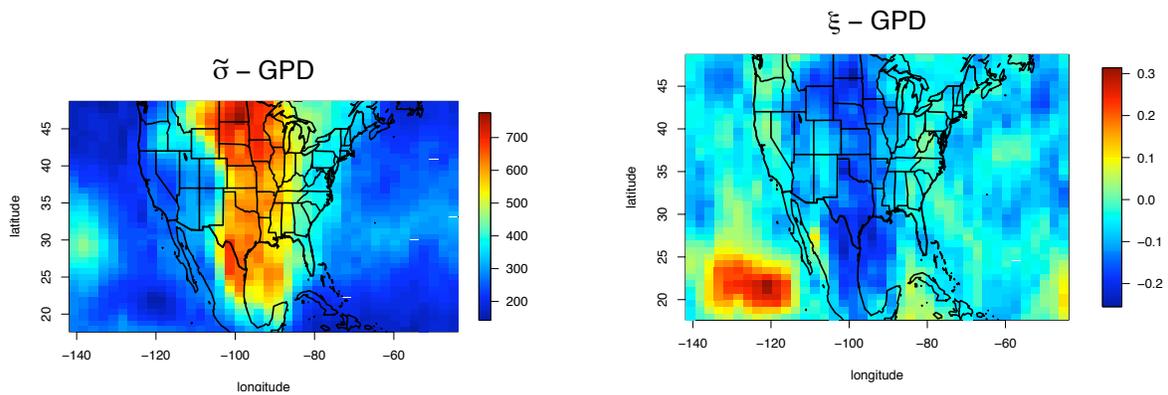


Figure 7: Heat Maps of the MLEs of the GPD Distribution Parameters

The maximum likelihood estimates of the GPD distribution parameters  $\tilde{\sigma}$  and  $\xi$  are shown in Figure 7. As in the previous section, estimates of the scale parameter (left panel) are comparatively large in the tornado alley region, indicating a higher variability of the data in that area. The estimated values of the shape

parameter  $\xi$  (right panel) are mostly negative, indicating a distribution that is bounded above. In the tornado alley, the values range from about -0.1 to approximately -0.24 in tornado alley.

The plug-in estimates of the 20-year return levels in Figure 8 are again comparatively high in tornado alley.

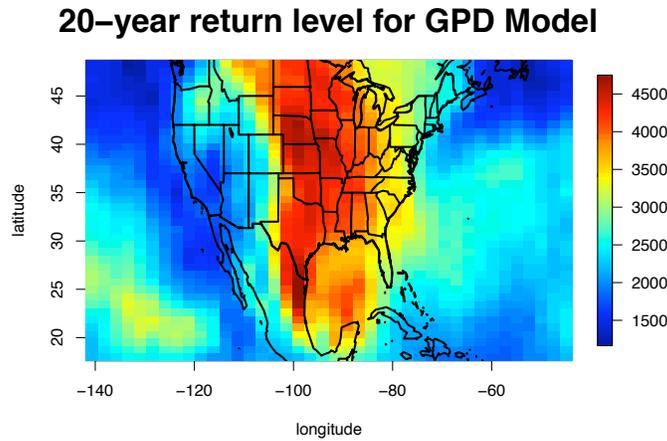


Figure 8: Heat Maps of the 20-Year Return Levels of the GPD Distribution

Finally, we assess goodness of fit in the same way as we did in the previous section on the GEV distribution. Figure 9 shows Q-Q plots and histograms for the same two locations. The next section features a comparison of the two approaches.

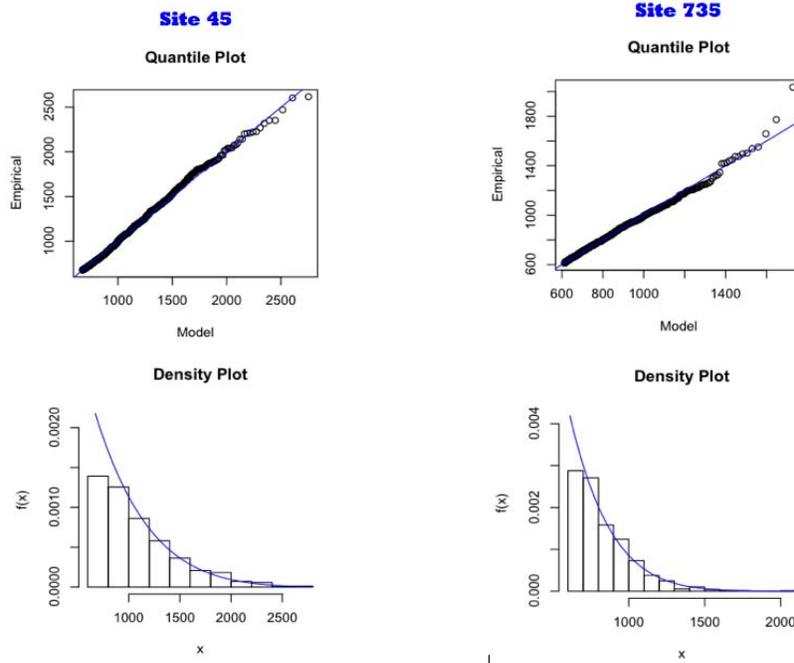


Figure 9: Q-Q and Histograms of sites 45 and 735 for GPD

### 3.2.3 Comparison of GEV and GPD

This section will compare different aspects of the GEV and the GPD method, and provides some visual comparisons of the results. Theoretically the  $\xi$  parameter is the same in the GEV and the GPD. However, due to different amounts of data used in the annual max analysis vs. exceedances, the parameter estimates

are different. The GEV only considers annual maxima, while the GPD tries to model all values that exceed a certain threshold  $u$ .

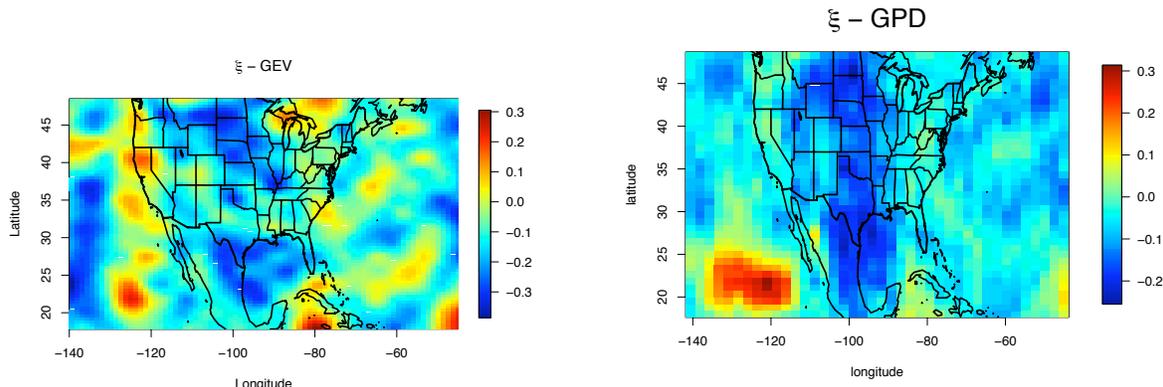


Figure 10: Maximum Likelihood Estimates of  $\xi$  for GEV (left) and GPD (right)

First we will compare the estimates of  $\xi$  plots between the two methods. In theory, these parameters are asymptotically equivalent. Of course, for any given data set, they will never be exactly the same for the two approaches. Figure 10 shows the two previously included maps of parameter estimates side by side for ease of comparison. These plots do not appear to be very similar. Dependences between subsequent exceedances in the GPD model did not seem to be responsible for the difference in the estimates, since declustering (see section 3.3) did not solve the issue. The discrepancies seem instead to be due to the fact that the GPD uses a much higher proportion of the original data values. Using annual maxima ( $1/365 = 0.274\%$  of the data) as in the GEV approach corresponds to using a threshold of approximately 99.7%. We found that after applying the same 99.7% threshold for the GPD, the  $\xi$  estimates are comparable.

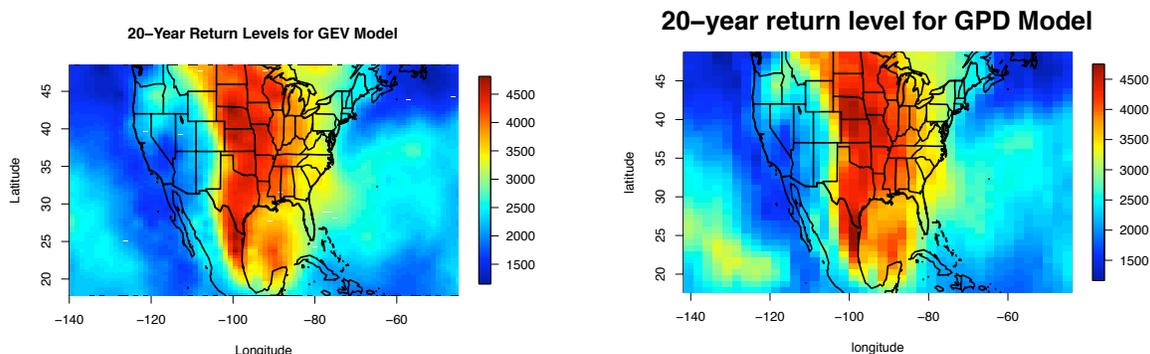
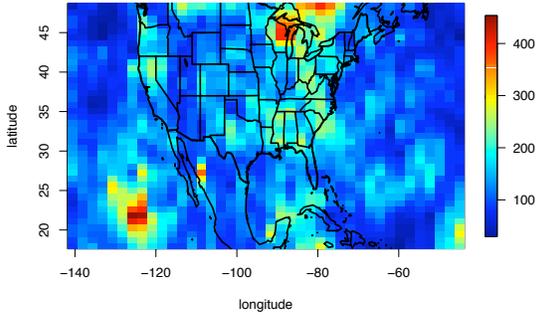


Figure 11: Return Level plots for GEV (left) and GPD (right)

Next, we compare our estimates for the 20-year return levels from the two approaches. The two heat maps in Figure 11 look very similar. This is reassuring, as different estimates for  $\xi$  in Figure 10 did not result in a large difference for the relevant statistic.

standard error for 20-year return level GEV



standard error for 20-year return level GPD

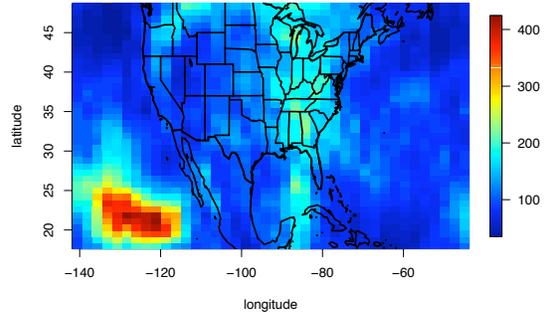


Figure 12: Standard Error of the 20-Year Return Levels for the GEV and GPD

We also examine the variability in the 20-year return level estimates from the two methods. Figure 12 shows plots of the standard error of the 20-year return level for the GEV and the GPD in the left and right panel, respectively. We see that the standard errors for the GPD appear to be lower than those for the GEV. This is to be expected, as more data (and therefore information) is used in the GPD approach.

As a last step, we put the diagnostic plots for the two approaches side-by-side. We see that both models fit the data fairly well, but the GPD appears to do a better job at the first site, which exhibited less extreme values. Overall, however, the goodness of fit is fairly similar.

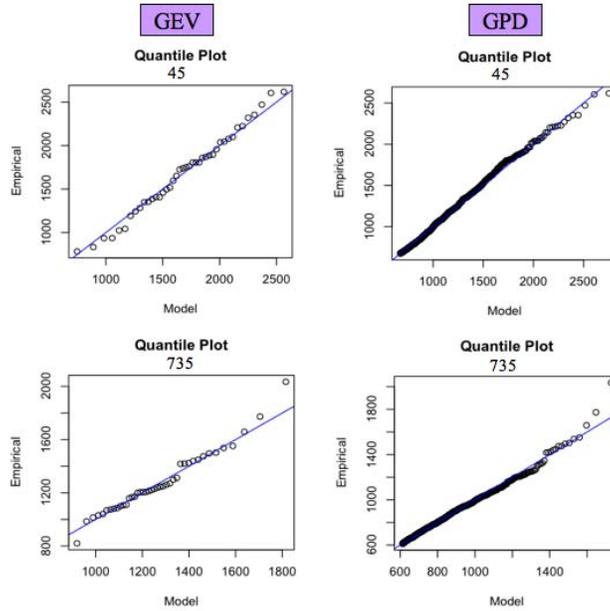


Figure 13: Q-Q Plots of 2 Stations for GEV (left) and GPD (right)

### 3.3 Declustering

When modeling exceedances using the GPD approach, it is often noted that subsequent extreme values are correlated over time, and tend to occur in clusters. This, of course, violates the assumption of independent observations, and might result in an inaccurate representation of the distribution of extreme values. As a simple and somewhat crude solution to this problem, a technique called runs declustering is often employed. To apply this technique, one must first choose a value  $r \geq 1$ . Two subsequent exceedances are then considered to be part of different clusters if there are at least  $r$  observations between them which do not exceed the threshold under consideration. Then, declustering simply discards all observations but the maximum for any

given cluster [14].

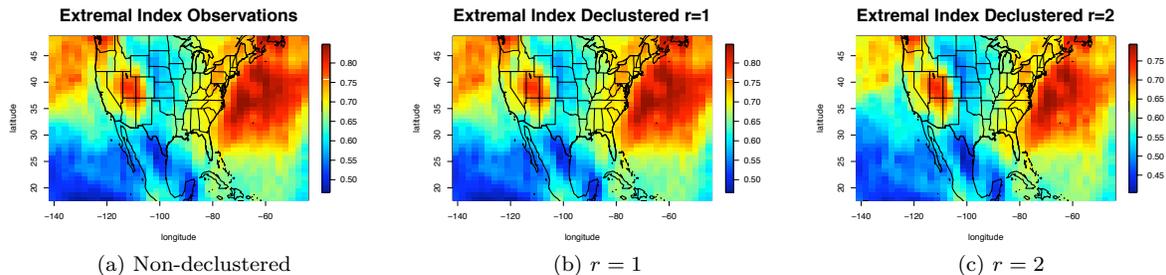


Figure 14: Extremal indices for all exceedances (left), after declustering with  $r = 1$  (middle), and after declustering with  $r = 2$  (right)

To assess the autocorrelation between threshold exceedances, one often considers a statistic called the extremal index [4]. The extremal index is bounded between 0 and 1, with large values indicating complete independence, and values near zero an indication of perfect dependence between the exceedances.

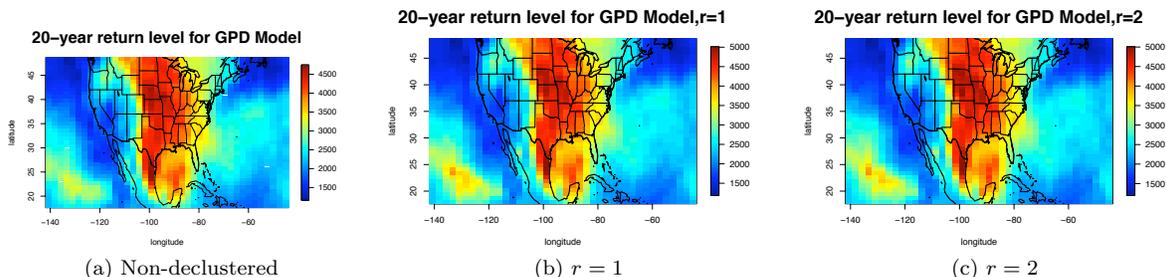


Figure 15: Return level for all exceedances (left), after declustering with  $r = 1$  (middle), and after declustering with  $r = 2$  (right)

We ran two declustering algorithms on our data, using  $r = 1$  and  $r = 2$ . The resulting extremal indices are shown in Figure 14. Ultimately we are interested in estimating return levels after fitting our models. Figure 15 shows return levels based on the GPD model using original and declustered data. We see from the two figures that, for our data set, declustering neither results in a big change in the extremal indices nor in the estimated return levels. Therefore, we did not explore this avenue of data analysis any further.

### 3.4 Time Trend

One of the main points of our analysis was to find out whether there was any indication that the extreme behavior of  $WmSh$  had changed over the 42-year time period under consideration. A change in scale and/or frequency of extreme values of  $WmSh$  would indicate a similar change in extreme weather events.

To explore this question, we included a time component in the GEV model. The model is very similar to the one in section 2.1, but now we assume that

$$\mu = \mu_0 + year * \mu_1.$$

After fitting the model to all 884 locations separately, we obtain a maximum likelihood estimate of  $\mu_1$  at each location. Figure 16 (a) shows that the estimated slope  $\mu_1$  is positive at many locations, thus indicating a shift of the location of the annual maxima towards higher values. The average of  $\mu_1$ , taken over all locations, is 216.9.

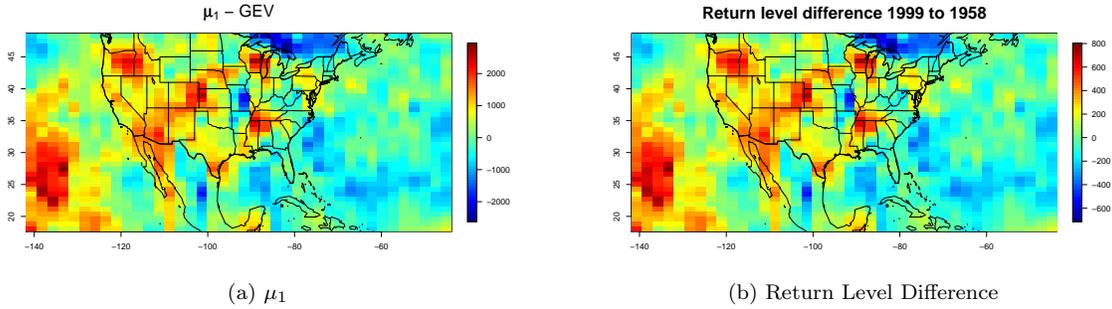


Figure 16: Heat maps of the slope coefficient over time for  $\mu$  (a), and the difference in estimated return levels between 1999 and 1958 (b)

Using these estimates, it is now possible to calculate return levels for any year within the time period of interest. In Figure 16 (b) we show the difference between the 20-year return levels for the year 1999, and the 20-year return levels for the year 1958. As expected, we see a similar behavior in these differences as we did for  $\mu_1$ . Many locations show an increase in the return levels, especially over the United States. The average difference, taken over all locations, is 59.0. This means that we estimate an average increase in *WmSh* of 59.0 in the region of interest from 1958 to 1999, based on the crude model fitted.

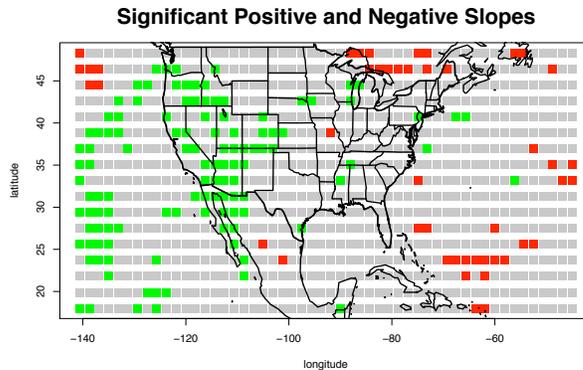


Figure 17: Locations with significant positive (green) and significant negative (red) slopes over time

This raises the question of whether the observed changes over time at some locations are just due to random variability, or whether they are due to a real change in the underlying processes. To answer this question, we calculated likelihood ratio test statistics to test the hypothesis “ $\mu_1 = 0$ ” for each location separately. Without accounting for multiple comparisons, of the 884 locations, 115 exhibited a statistically significant positive slope, while 46 exhibited a significant negative slope (all at the  $\alpha = 0.05$  level). Figure 17 shows the locations of the locations where the null hypothesis  $\mu_1 = 0$  was rejected.

To actually control the family-wise error rate at  $\alpha = 0.05$  when conducting multiple statistical tests, one must account for the fact that more than one test is carried out simultaneously. We decided to use the sequential step-down procedure introduced by Holm [13] to adjust our p-values. Using this procedure, only four locations showed a significant trend over time, all of them positive.

## 4 Analysis using Bayesian Spatial Extreme Value Models

Based on the findings of our exploratory data analysis, we fit two Bayesian hierarchical models using the GPD as a likelihood function. Alternatively, we could have used the GEV distribution as a likelihood as in, e.g. [16], but by using the GPD we exploit a larger amount of data. Both models allow parameter values to vary spatially but the two models differ in their treatment of spatial association. By fitting each model

using Bayesian techniques, the uncertainty associated with parameters and return-levels under each model is properly accounted for. Additionally, the Bayesian framework provides sufficient flexibility to fit large and complicated models. Markov chain Monte Carlo (MCMC) methods are employed to obtain a sample of size 5,000 from the posterior distribution after an initial burn-in period of 5,000 iterations. Four parallel chains were used to assess convergence.

## 4.1 Model 1

The first, and most basic, Bayesian model we considered is based on a GPD model (see section 2.2) for the exceedances  $y_k(\mathbf{s}_i) = x_k(\mathbf{s}_i) - u(\mathbf{s}_i)$  with different shape and scale parameters for tornado alley and outside of tornado alley. Specifically, consider the model defined by,

$$\begin{aligned}
 y_k(\mathbf{s}_i) | y_k(\mathbf{s}_i) > 0, \sigma(\mathbf{s}_i), \xi(\mathbf{s}_i) &\sim GPD(\sigma(\mathbf{s}_i), \xi(\mathbf{s}_i)), \\
 k &= 1, \dots, K_i \\
 i &= 1, \dots, n \\
 u(\mathbf{s}_i) &= 95\text{th percentile of } \mathbf{x}(\mathbf{s}_i) \\
 \sigma(\mathbf{s}) &= \sigma + \alpha_\sigma \mathbf{1}_{\mathbf{s} \in \mathcal{A}_\sigma} \\
 \xi(\mathbf{s}) &= \xi + \alpha_\xi \mathbf{1}_{\mathbf{s} \in \mathcal{A}_\xi}
 \end{aligned} \tag{4}$$

The sets  $\mathcal{A}_\sigma$  and  $\mathcal{A}_\xi$  refer to arbitrarily defined regions which roughly align with tornado alley. We defined the regions based on the heat maps of GPD MLEs (Figure 7). Alternatively, we could estimate  $\mathcal{A}$  via, e.g. a discrete mixture of normals, but manually defining tornado alley in this fashion is sufficient given the research questions considered in this report.

To complete the model specification, the prior distributions for the unknown parameters are assumed to be,

$$\begin{aligned}
 \sigma &\sim IG(2.1, 500) \\
 \alpha_\sigma &\sim N(0, 100^2) \\
 \xi &\sim N(0, .1^2) \\
 \alpha_\xi &\sim N(0, .1^2),
 \end{aligned}$$

where each parameter is assumed *a priori* independent. As a general rule, we tried to use prior distributions which were relatively diffuse but still place high mass on reasonable parameter values. However, given the size of the data set, any prior will essentially be overwhelmed by the likelihood. In contrast, [5] describes an approach for eliciting informative prior distributions but the prior distributions given above worked well for our models.

The posterior means for the parameters were found to be  $\hat{\sigma} = 265.07$ ,  $\hat{\alpha}_\sigma = 281.46$ ,  $\hat{\xi} = 0.04$ , and  $\hat{\alpha}_\xi = -0.09$ . Figure 18 displays the 20-year return levels over the region of interest.

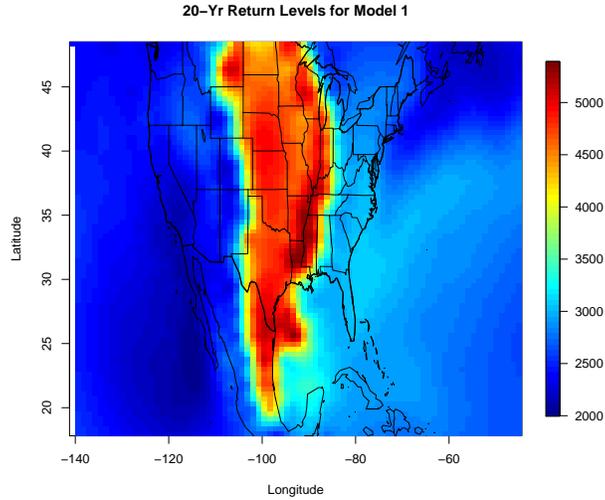


Figure 18: Posterior mean of 20-year return levels for Model 1.

As expected, Model 1 showed that values of  $WmSh$  were higher for the tornado alley region. However, contrasting Figure 18 with Figure 11 shows that Model 1 is, perhaps, more simple than the process displayed by the data.

## 4.2 Model 2

As an extension of Model 1, we allow  $\sigma(\mathbf{s}_i)$  to vary across locations by assuming

$$\log(\sigma(\mathbf{s}_i)) = w(\mathbf{s}_i) \sim GP(\mu_w, \tau^2 \rho_\phi(\cdot))$$

where  $GP(\cdot)$  is a Gaussian process and  $\rho_\phi(\cdot)$  is the exponential correlation function such that

$$Cov(w(\mathbf{s}_i), w(\mathbf{s}_j)) = \tau^2 \exp\{-\phi \times \|s_i - s_j\|\}$$

and  $\|\cdot\|$  is spherical distance. Even though examination of Figure 7 gives evidence towards the use of a non-stationary process, the methodology to implement such a process is beyond the scope of this report so the simpler stationary Gaussian process is assumed. The prior distributions for  $\xi$  and  $\alpha_\xi$  given above in Section 4.1 are again used for Model 2. Additionally, we assume  $\phi \sim IG(3, .3)$ ,  $\tau^2 \sim IG(2.1, 1)$ , and  $\mu \sim Unif(-\infty, \infty)$  as prior distributions for the Gaussian process prior for  $w(\mathbf{s}_i)$ .

Again using 5,000 draws from the posterior distribution, the posterior means of the parameters are as follows:  $\hat{\xi} = -0.02$ ,  $\hat{\alpha}_\xi = -0.12$ ,  $\hat{\mu} = 5.52$ ,  $\hat{\phi} = 0.01$ , and  $\hat{\tau}^2 = 0.05$ . Of particular interest is the difference in the estimates of  $\xi$  and  $\alpha_\xi$  between Model 1 and Model 2. Specifically, Model 2 estimated  $\xi$  to be negative compared to a positive estimate under Model 1. This could possibly be due to the fact that a more realistic model for  $\sigma(\mathbf{s})$  was used. Thus, because more of the variability in the data was explained by  $\sigma(\mathbf{s}_i)$ ,  $\hat{\xi}$  was lower.

Figure 19 displays the posterior mean of the 20-year return level as estimated using Model 2. Notice that Model 2 estimates slightly smaller 20-year return values over the region of interest. This is due to the lower (negative) estimate of  $\xi$  as we have discussed above. These lower return values, however, are seen to be more realistic return values than those given by Model 1 as Model 2 is a superior fit to the data than Model 1.

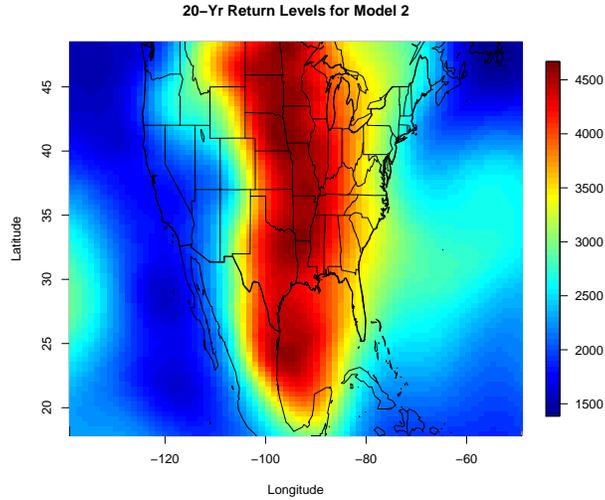


Figure 19: Posterior mean of 20-year return levels for Model 2.

### 4.3 Model Comparison

A natural question is whether or not the added complexity of Model 2 is necessary to obtain accurate and precise predictions of extreme weather. To investigate this further, Model 1 and Model 2 were compared using various model fit diagnostics. The deviance information criterion (DIC) is a model selection tool (similar to AIC or BIC) which measures overall model fit, for more information refer to [19]. Lower values of DIC correspond to better fit models. According to the values of DIC displayed in Table 2, the additional complexity of Model 2 is beneficial as it increases model fit (i.e. decreases DIC).

To compare Model 1 and Model 2 further, the root mean square error,

$$\text{RMSE} = \sqrt{\sum_i \sum_k (y_k(\mathbf{s}_i) - \tilde{y}_k(\mathbf{s}_i))^2},$$

where  $\tilde{y}_k(\mathbf{s}_i) = \mathbb{E}(y_k(\mathbf{s}_i)|y)$  and  $y$  is all the data, the coverage probability of a 95% predictive interval for  $y_k(\mathbf{s}_i)$ , and the average length of the predictive intervals for observations inside and outside tornado alley were calculated and are presented in Table 2. Note that here, *CI* stands for *credible interval*. As Table 2 displays, we found that by exploiting the spatial correlation in  $\sigma(\mathbf{s}_i)$  we were able to decrease the amount of uncertainty associated with extreme events.

Table 2: Model fit diagnostics for Model 1 and Model 2.

Diagnostic	MLE (GPD)	Model 1	Model 2
DIC	NA	2814.5	2724.8
RMSE	394.47	323.92	313.31
Coverage	0.95	0.94	0.95
Avg. Width of CI (Alley)	1294.19	1046.10	584.96
Avg. Width of CI (Other)	1368.19	1180.42	961.30

### 4.4 Analysis of Temporal Trend

Having developed a spatial model, we turn to the question of the temporal trend in extreme events. The temporal aspect of extreme events is not explicitly included in Model 2 but in Section 5 we develop a spatio-

temporal model for  $WmSh$  based on a point process for extreme events. Rather, here we answer the question of a temporal trend by dividing the data set into two pieces: pre-1972 and post-1985. Data between 1973 and 1984 were discarded for this section. Model 2 was fit to both data sets and the differences in the posterior mean of the 20-year return levels are displayed in Figure 20. While Figure 20 displays an increase in the intensity of extreme storms, the increase is only slight. The largest increase in  $WmSh$  was in the South Pacific which increased by 660.23. Again, most of the region under consideration showed an increase in  $WmSh$  but further analysis is needed to determine whether or not this increase is practically significant (e.g. can cause significant additional damage).

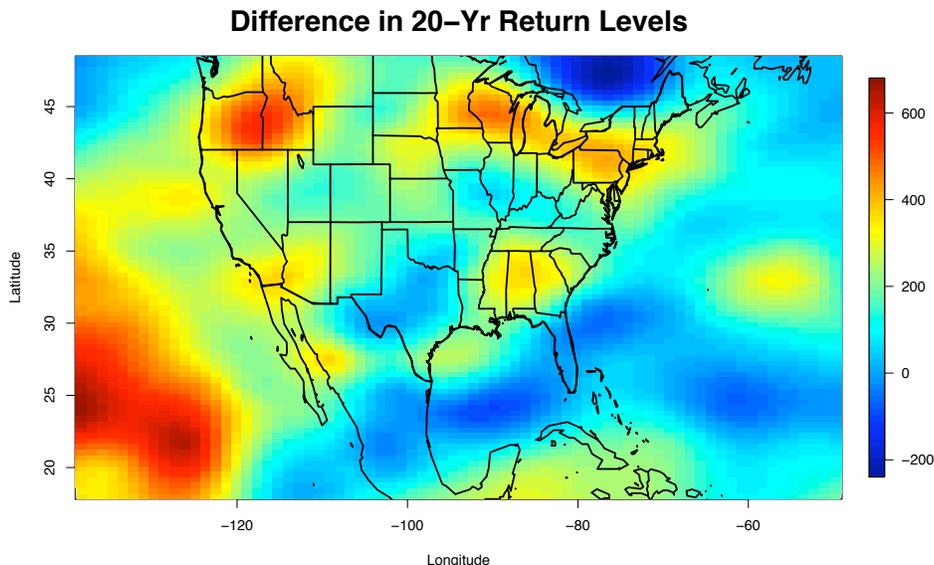


Figure 20: Difference (1999 minus 1974) in posterior mean of 20-year return levels under Model 2.

## 5 Conclusion and Future Work

The purpose of this report was to model the large-scale indicator  $WmSh$  as a predictor of extreme weather. Because  $WmSh$  can be measured and predicted in real time up to 3 hours in advance, residents can be warned early of an approaching storm. Specifically, this report worked with daily re-analysis data during the years 1958 through 1999 from 884 locations across the continental United States and Mexico. As we were mostly interested in extreme weather, we used the generalized extreme value and generalized Pareto distributions to model the tail behavior of  $WmSh$  so as to properly predict extreme weather.

First, using maximum likelihood, we considered each location independently and obtained estimates of the parameters of the GEV and GPD. Consistent with prior knowledge regarding locations of tornadoes, we found distinctly larger return levels for the mid-western region of the United States near Kansas and Oklahoma (see Figure 11). Additionally, we found strong spatial patterns in the estimates of the parameters. To account for this spatial behavior, we developed, fit, and compared two hierarchical models using Markov chain Monte Carlo techniques. The second hierarchical model, which accounted for spatial correlation through the use of a Gaussian process prior on the scale parameter of the GPD, was found to be superior to the simpler model by improving predictions and decreasing the uncertainty associated with extreme weather.

While a suitable spatial model was found which improves predictions of extreme weather, more advanced spatio-temporal models could be developed which further improve predictions and decrease uncertainty. We are continuing to develop a model, similar to the model of [6], based on a point process approach for extremes

(see [17]). Specifically, using a point process approach, the likelihood for  $\{x_k(\mathbf{s}_i)|x_k(\mathbf{s}_i) > u(\mathbf{s}_i)\}$  is given by,

$$\prod_i \prod_k \exp \left\{ - \int_0^T \int_{u(\mathbf{s}_i)}^{\infty} \lambda(k, x_k(\mathbf{s}_i)) dx dk \right\} \lambda(k, x_k(\mathbf{s}_i)),$$

where,

$$\lambda(t, y) = \frac{1}{\sigma_t(\mathbf{s})} \left[ 1 + \xi_t(\mathbf{s}) \left( \frac{y - \mu_t(\mathbf{s})}{\sigma_t(\mathbf{s})} \right) \right]^{-\frac{1}{\xi_t(\mathbf{s})} - 1},$$

and  $(\mu, \sigma, \xi)$  are the corresponding parameters of the GEV distribution. Specifically we propose the model,

$$\begin{aligned} \mu_t(\mathbf{s}) &= \beta_0 + \tilde{\beta}_0(\mathbf{s}) + \beta_1 \times t + \tilde{\beta}_1(\mathbf{s}) \times t \\ \xi_t(\mathbf{s}) &= \xi + \alpha_\xi \mathbf{1}_{\mathbf{s} \in \mathcal{A}_\xi} \\ \log(\sigma_t(\mathbf{s})) &= \mu_\sigma + \omega(\mathbf{s}), \end{aligned} \tag{5}$$

where  $(\tilde{\beta}_0(\mathbf{s}), \tilde{\beta}_1(\mathbf{s}), \omega(\mathbf{s}))$  follow a multivariate mean zero Gaussian process induced via coregionalization (see [9]). This model would allow a linear trend in time while account for the underlying spatial structure of *WmSh* and should provide an improvement over the models developed in this report. Due to time constraints on the current project, however, we were unsuccessful in coding an MCMC algorithm to obtain posterior draws from this model.

Throughout this report we only considered modeling *WmSh*. Perhaps a better approach would be to jointly model *CAPE* and *WS* as a bivariate extreme. Developing a bivariate model has several challenges in that, while the theory underlying bivariate extremes is well-developed, standard approaches to bivariate modeling are less developed. Additionally, the increase in dimensionality presents new challenges for model selection, fitting, and validation. There is a final concern with bivariate models in that the strength of dependence between two variables can, perhaps, diminish at the extreme levels. We feel that this final concern is not the case for the re-analysis data considered here and a bivariate approach may prove useful in predicting and modeling extreme weather.

Perhaps most importantly, we have discussed relatively little about the relationship between *WmSh* and tornadoes and hurricanes. While high values of *WmSh* are typically associated with extreme weather, further work is needed to determine the levels of *WmSh* which are associated with severe weather. While [1] has addressed this question, we feel that this question is still open to further research and inquiry.

## References

- [1] Harold E. Brooks, James W. Lee, and Jeffrey P. Craven. The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmospheric Research*, 67-68:73–94, July - September 2003.
- [2] Fortran code by H. Akima R port by Albrecht Gebhardt aspline function by Thomas Petzoldt enhancements and corrections by Martin Maechler. *akima: Interpolation of irregularly spaced data*, 2009. R package version 0.5-2.
- [3] Original S code by Richard A. Becker and Allan R. Wilks. R version by Ray Brownrigg Enhancements by Thomas P Minka. *maps: Draw Geographical Maps*, 2009. R package version 2.1-0.
- [4] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag London Limited 2001, London, Great Britain, 2004.
- [5] Stuart G. Coles and Jonathan A. Tawn. A bayesian analysis of extreme rainfall data. *Applied Statistics*, 45(4):463–478, August 1996.
- [6] Dan Cooley, Doug Nychka, and Philippe Naveau. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840, September 2007.

- [7] Original S functions by Stuart Coles, R port, and R documentation files by Alec Stephenson. *ismev: An Introduction to Statistical Modeling of Extreme Values*, 2009. R package version 1.33.
- [8] Reinhard Furrer, Douglas Nychka, and Stephen Sain. *fields: Tools for spatial data*, 2009. R package version 5.02.
- [9] A. E. Gelfand, A. M. Schmidt, S. Banerjee, and C.F. Sirmans. Nonstationary multivariate process modeling through spacially varying coregionalization (with discussion). *Test*, 13(2):1 – 50, 2004.
- [10] Eric Gilleland. Statistical software for weather and climate. Presentation, March 2009.
- [11] Eric Gilleland, Rich Katz, and Greg Young. *extRemes: Extreme value toolkit*, 2009. R package version 1.60.
- [12] Eric Gilleland, Matthew Pocerlich, Harold E. Brooks, and Barbara G. Brown. Large-scale indicators for severe weather. Manuscript in Preparation, 2008.
- [13] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.
- [14] Elizabeth Mannshardt-Shamseldin. Severe weather under a changing climate: large scale indicators of extreme events. IMSM Presentation, July 2009.
- [15] M. Pocerlich, E. Gilliland, H. Brooks, B. Brown, and P. Marsh. Analysis of atmospheric conditions conducive to small scale extreme events from larger scale global reanalysis data. Manuscript in Preparation, 2008.
- [16] Huiyan Sang and Alan E. Gelfand. Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics*, 16(3):407–426, September 2009.
- [17] Richard L. Smith. Extreme value analysis of environmental time series: an example based on ozone data (with discussion). *Statistical Science*, 4:367–393, 1989.
- [18] Richard L. Smith. Statistics of extremes, with applications in environment, insurance, and finance. Department of Statistics, University of North Carolina, Chapel Hill, NC, USA, March 2003.
- [19] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, 64(4):583–639, October 2002.
- [20] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

# Stress Tests of Risk Bearing Enterprises

Robertas Gabrys<sup>1</sup>, Jenya Kirshtein<sup>2</sup>, Nitseh Kumar<sup>3</sup>, Hongxia Yang<sup>4</sup>, Kun Zhou<sup>5</sup>, James Zou<sup>6</sup>

Faculty Mentor:  
Dr. Jeff Scroggs  
North Carolina State University

Problem Presenter:  
Dr. Frank Mayer  
Republic Mortgage Insurance Company

## Abstract

We propose a methodology to explore the distribution of the cumulative cash flow generated by a portfolio of insured mortgage loans for various dependencies among geographic regions. We use historic house price indices, interest rates, defaults and prepayment data to predict the future house price indices, defaults and prepayments which in turn help us calculate the cumulative cash flow. The methodology can be relatively easily computed using MATLAB.

## 1 Preliminaries

We consider the data set containing information about the loans issued between 1980 and 1997 in 50 US states. The data shows how many loans change status in each of 12 consecutive years from origination date in each region. The following statuses can occur during the lifetime of a loan:

- *PSTND*, *Persistent Never-Delinquent*, loan is in progress and borrower never misses a payment. All loan policies start in this status.
- *NDROFF*, *Never-Delinquent Runoff*, borrower never misses a payment but decides to refinance and the loan therefore terminates. We consider it to be the case of *prepayment*.
- *PSTED*, *Persistent Ever-Delinquent*, loan gets this status as soon as the borrower ever misses a payment. After that, the loan can either stay in progress in less reliable status *or* become *EDROFF*.
- *EDROFF*, *Ever-Delinquent Runoff*, borrower terminates the loan while having a missed payment on record. We consider it to be the case of *default*. *EDROFF* can have two outcomes,  $EDROFF=EDNC+TC$ .
- *EDNC*, *Ever-Delinquent Non-Claim*, borrower terminates the loan, bank doesn't file the claim to the insurance company since the bank's costs can be compensated from other sources (e.g. selling the house for at least the amount of the loan balance or borrower paying the loan in full).
- *TC*, *Termination with Claim*, borrower terminates the loan, bank files the claim to the insurance company.

---

<sup>1</sup>Utah State University

<sup>2</sup>University of Denver

<sup>3</sup>University of California, Merced

<sup>4</sup>Duke University

<sup>5</sup>Pennsylvania State University

<sup>6</sup>Harvard University

**Income Stream** The only inflow of cash for an insurance company is the fixed premium charged by the company for a policy. The company suffers an outflow of cash if a policy defaults for which the company has to bear a loss. Another term which goes into the calculation of income stream is the number of policies that stop paying the premium because they settle the loan by prepayment. This term freezes the income flow.

Lets define a cohort as different policies originating in the same year. In the synthetic data provided to us, the loan amount the number of policies are fixed for every cohort. So, for each cohort, let us define the following terms:

- $d_t$  = number of defaults at time  $t$
- $g$  = optional guarantee
- $I$  = Installment paid towards the loan/period
- $b_t$  = balance at time  $t$  of the loan amount still to be paid
- $l_t$  = loss due to defaults at time  $t$
- $n_t$  = total number of policies in each cohort at time  $t$
- $p_t$  = number of prepayments at time  $t$
- $L$  = loan amount per policy
- $i_t$  = net income at time  $t$
- $P$  = fixed premium/period

The balance at time  $t$  is calculated by  $b_t = L - I(t - 1)$  and the loss due to defaults is then calculated using:

$$l_t = d_t b_t g = d_t I(t - 1)g, \text{ and,} \quad (1)$$

$n_t = n - d_t - p_t$ , where  $n$  is the initial number of policies in each cohort.

$$i_t = P n_t - l_t \quad (2)$$

Let  $N$  be the total number of loans issued in each region (assume the same number of loans in each state). We calculate the default and prepayment probabilities for each year of the lifetime of a loan in each region. E.g., given the number of loans that default ( $d_i$ ) and prepay ( $p_i$ ) in year  $i$  ( $i = 1 \dots 12$ ) from fixed origination date in a fixed region, we calculate the probability of default  $D_i$  and prepayment  $P_i$  in year  $i$  as follows:

$$D_i = \frac{d_i}{N - \sum_{j=1}^{i-1} d_j - \sum_{j=1}^{i-1} p_j}$$

$$P_i = \frac{p_i}{N - \sum_{j=1}^{i-1} d_j - \sum_{j=1}^{i-1} p_j}$$

## 2 Simulations and applications

We perform a simulation study to investigate how our proposed method works on simulated data. We model the time variation of interest rate  $r$  and housing index  $\mathbf{H}_i, i \in \{1, 2, \dots, 9\}$  as Stochastic Differential Equations (SDE):

$$dr(t) = (r_0 - r(t))dt + \sigma r^\alpha W_r(t) \quad (3)$$

$$d\mathbf{H}_i(t) = \bar{\mathbf{H}}_i(t) + \Sigma d\mathbf{W}_i(t), \quad i = 1, \dots, 9, \quad (4)$$

where  $r_0$  is the expected mean interest rate over years (set at 0.05 for now) and  $W_r$  and  $\mathbf{W}_i$  are Brownian motions representing fluctuations of the indices of treasury bills and house price market index correspondingly. Note that matrices are in bold. We chose these equations to be consistent with the idea that interest rate should not have any drift (i.e. mean reverting) and that the housing index should have a drift comparable to

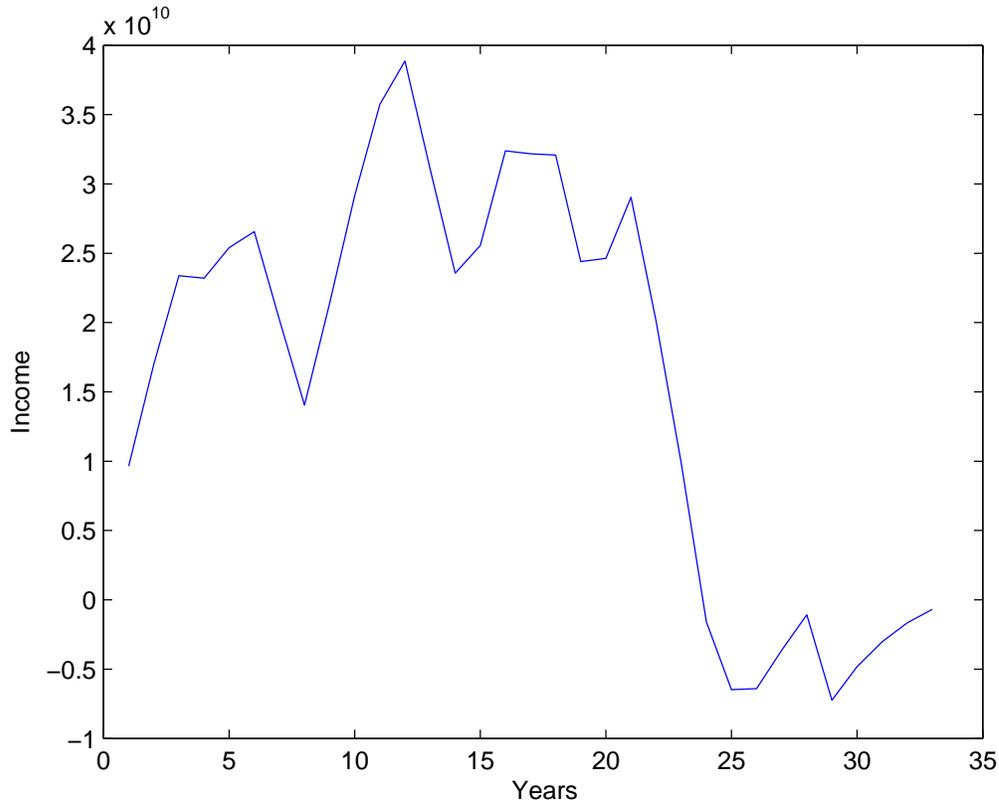


Figure 1: Historic Income Stream.

inflation. We also allow the possibility that housing indices in different regions are correlated with each other. To simulate "dependent" multivariate Brownian bridges  $\mathbf{W}_i$  we sample from a joint multi-variate Gaussian distribution with known covariance matrix. More precisely, suppose there are 9 regions. Choose  $\mu$  to be a 9 dimensional vector of means and  $\Sigma$  to be a  $9 \times 9$  covariance matrix, and let  $MVN(\mu, \Sigma)$  be the corresponding multivariate normal. The forward Euler solution of SDE works as follows:

1. At time  $t = 0$ , initialize  $\mathbf{H}$  to be a vector of initial housing indices.
2. Sample  $\xi \sim MVN(\mu, \Sigma)$ .
3. Set  $\mathbf{H}(\mathbf{t} + \Delta\mathbf{t}) = \mathbf{H}(\mathbf{t}) + \bar{\mathbf{H}}\Delta\mathbf{t} + \xi\sqrt{\Delta\mathbf{t}}$ .
4. Iterate steps 2 and 3 until reach final time  $T$ .

By varying the off diagonal entries of the covariance matrix we can asses the impact of the interconnections between different housing markets. Figure 3 displays housing indices in two correlated regions.

The SDEs are solved numerically using forward Euler's method.

Figure ?? illustrates interest returns simulated interest rates for the whole 20 year period. A sample run of the annual interest rate SDE gives the 20 year interest rate as follows  $r(t) = \{5.0049, 5.1396, 3.9439, 4.1021, 5.3583, 4.7933, 5.4302, 5.0709, 4.6832, 5.0007, 4.2242, 4.5853, 5.8883, 5.8684, 5.4519, 5.7423, 6.1391, 4.6700, 5.2552, 5.3029\}$ .

We assume the refinance (prepayment) probability  $P(t)$  is inversely proportional to the interest rate  $r(t)$ , i.e.  $P \propto 1/r$ . Since  $r(t)$  varies from year to year, the refinance probability also changes. Then using the simulated interest rates, we model refinance as a simple binomial model. For year  $t$ , with probability  $P(t)$ , a house refinances and hence stops an insurance payment. That is we expect  $P(1)$  fraction of the original houses

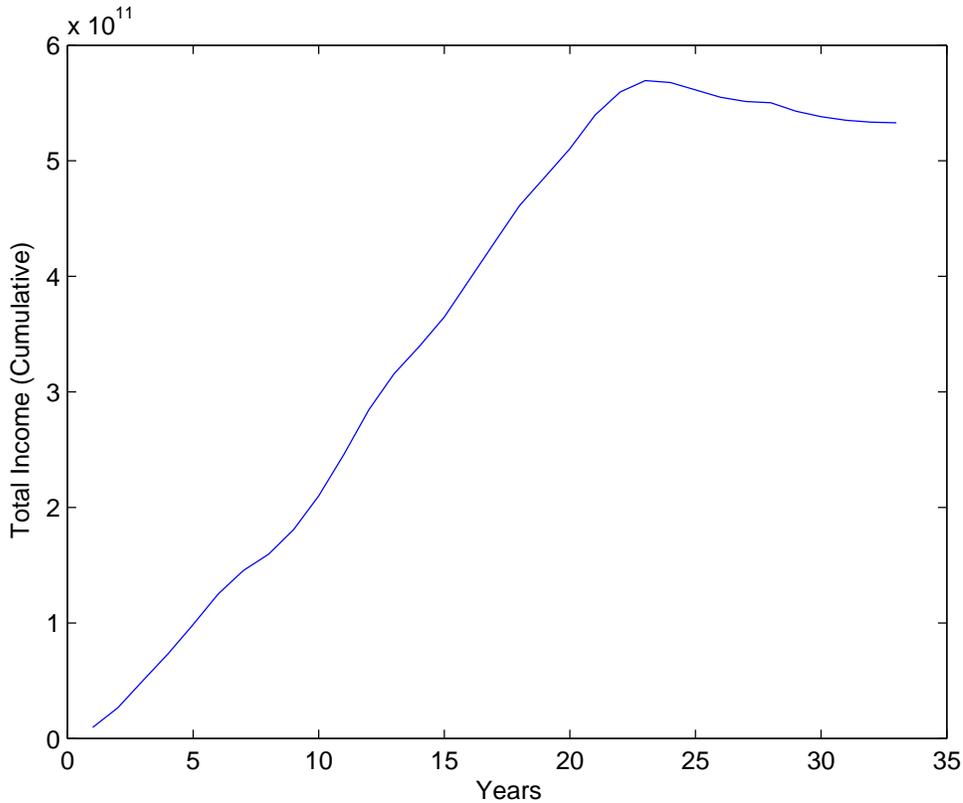


Figure 2: Cumulative Historic Income Stream.

to refinance after one year. If a house remains after one year, then with probability  $P(2)$  it refinances after the second year. And so on. If it does not refinance after 20 years, then it has paid its full insurance contract.

To investigate the distribution of the durations of the contracts till termination or end, we simulate a large number of houses and how long (in years) each house stays under a contract. Figure 4 displays durations till termination of the insurance contract and Figure 5 displays the histogram of those durations. For the simulation shown, the average number of years under contract is 14.12 and the variance is 47.56. The income stream from this cohort of houses is proportional to the aggregate number of years that these houses stay under contract.

### 3 Default and Prepayment Rate Modeling

In this section we introduce two different models to simulate the default and the prepayment rates. First methods deals with Bayesian approach and the second methods employs the linear regression with MLE.

Probability of prepayment at time  $t$  for the loan initiated at  $t_0$  at region  $i$ :

$$y_{t,t_0}^P = f(t - t_0)e^{\beta(r_t - r_{t_0}) + \gamma(h_t - h_{t_0})/h_{t_0}} \quad (5)$$

where  $t_0$  denotes the time of the origination of a loan,  $h_{it}$  is the house price at time  $t$  and region  $i$ , and

$$f(t - t_0) = \begin{cases} a(t - t_0), & t - t_0 \leq 3 \\ 3a, & \text{otherwise.} \end{cases} \quad (6)$$

Analogously the probability of default at time  $t$  for the loan initiated at  $t_0$  at region  $i$ :

$$y_{t,t_0}^D = f(t - t_0)e^{\gamma(h_t - h_{t_0})/h_{t_0}} \quad (7)$$

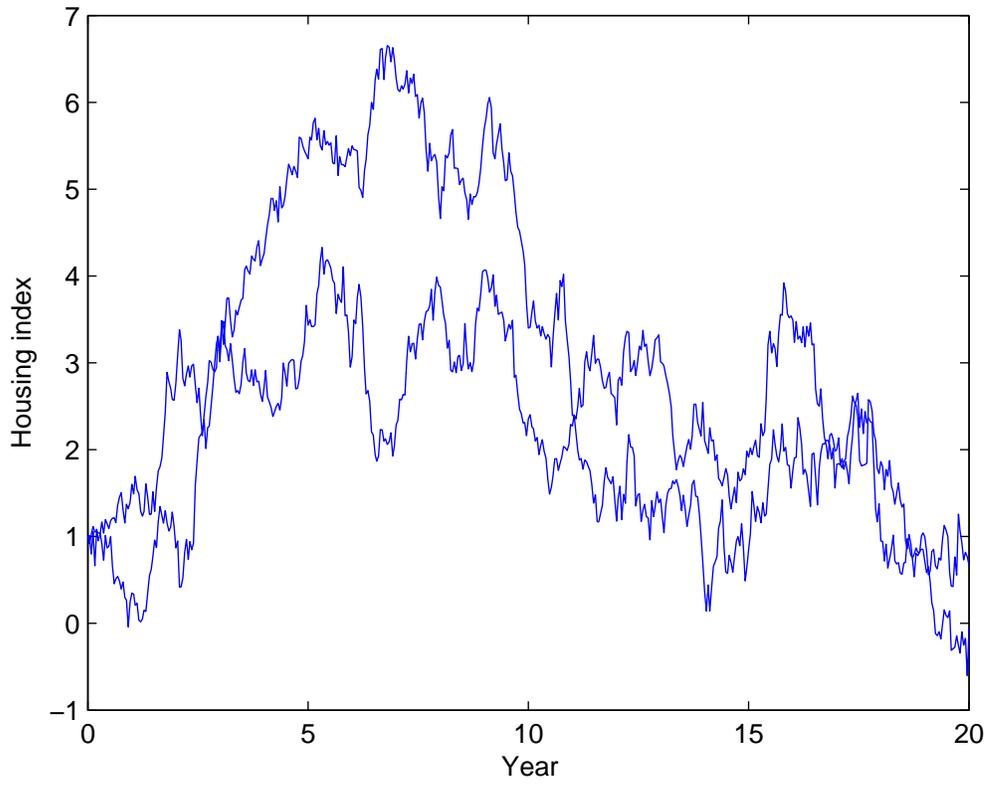


Figure 3: Two simulated housing indices

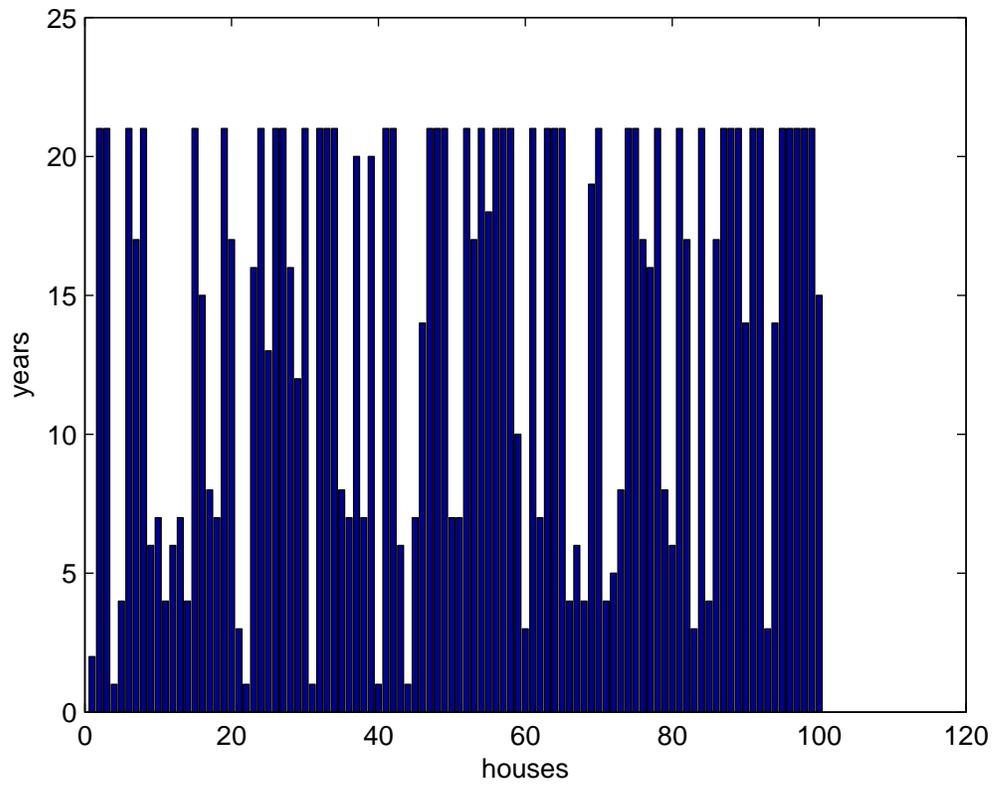


Figure 4: Simulated time durations till termination

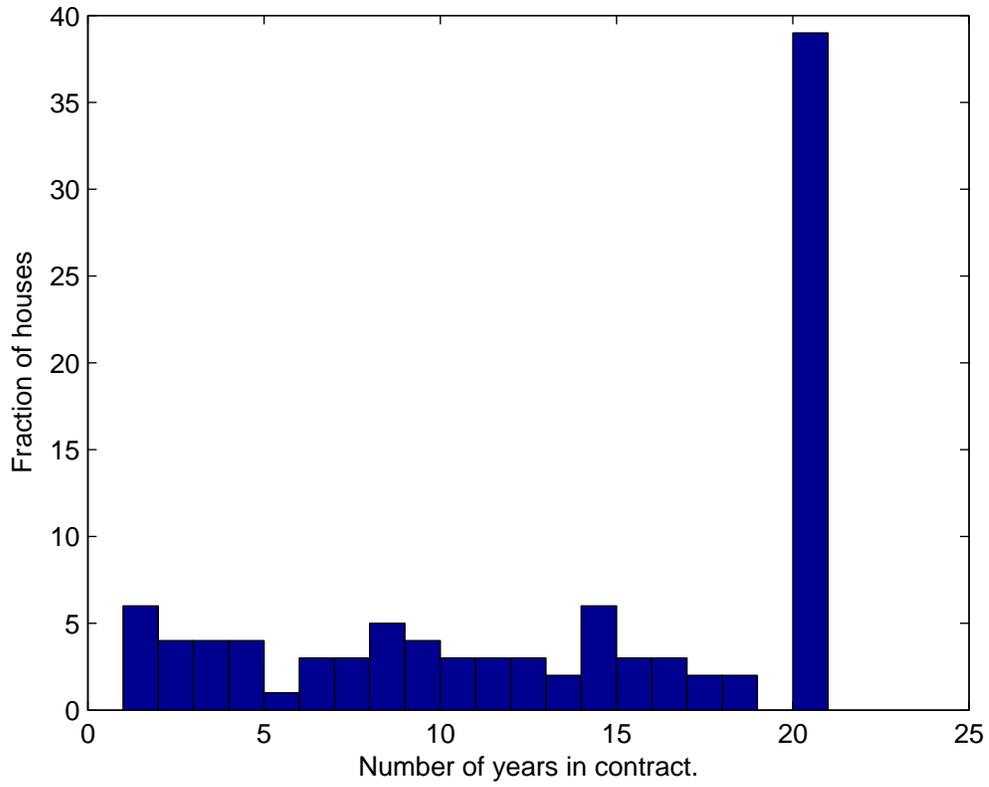


Figure 5: Histogram of simulated time durations till termination

### 3.1 Bayesian Parametric Model. Priors

In this section we will list the assumptions, i.e. prior distributions of the parameters in (??).

- For the sake of simplicity we assume that  $a(t-t_0) \sim N(\mu, \sigma^2)$  and  $a(t-t_0) \in (0, 1) \forall t$ . Taking logarithm of  $a(t-t_0)$ , implies that  $a^* \in (-\infty, 0)$ . Thus we truncate  $a^*$  at 0 and consider only negative values. Moreover, it is natural to assume  $a^* \sim N^-(\mu_{a^*}, \sigma_{a^*}^2)$ .
- Since  $r_t \in (0, 1)$ , we let  $r_t \sim Unif(0, 1)$ .
- Since  $h_{it} > 0$  (house price cannot be negative), we assume  $h_{it} \sim N^+(\mu_{h_{it}}, \sigma_{h_{it}}^2)$ .
- Let  $\epsilon_{it} \sim N(0, \sigma_{it}^2)$  is an idiosyncratic term allowing region  $i$  at time  $t$  to respond differently than expected based on their overall logarithm of default value. Let  $\delta_i \sim \Gamma(a, b)$ ,  $\gamma_t \sim \Gamma(\frac{1}{2}, \frac{1}{2})$  and  $a \sim \Gamma(2, 1)$ ,  $b \sim \Gamma(2, 1)$ . Suppose  $\sigma_{it}^{-2} = \delta_i \gamma_t$ , so that we have an additive log-linear model for the residual precision which will account for variance both for region  $v$  and time  $t$ , and we also assume for simplicity that there is no interaction. This allows the residual variance to change for the different regions ( $\delta_i$ ) and also allows to account for variability over time ( $\gamma_t$ ).
- For coefficients  $b$  and  $c$ , we assume that  $b \sim N(\mu_b, \sigma_b^2)$  and  $c \sim N(\mu_c, \sigma_c^2)$

#### 3.1.1 Gibbs Sampler

Next we use Gibbs sampler to update the prior distributions introduced in the previous section.

- Sample  $a^*$  from the truncated normal distribution, i.e.

$$a^* \sim N^- \left( \frac{\sum_i \sum_t \frac{1}{\sigma_{it}^2} (y_{it}^* - br_t - c(h_{it} - h_{it_0})) + \frac{1}{\sigma_{a^*}^2} \mu_{a^*}}{\sum_i \sum_t (\frac{1}{\sigma_{it}^2} + \frac{1}{\sigma_{a^*}^2})}, \frac{1}{\sum_i \sum_t (\frac{1}{\sigma_{it}^2} + \frac{1}{\sigma_{a^*}^2})} \right)$$

- Sample  $b$  from the following truncated normal distribution, i.e.

$$b \sim N^- \left( \frac{\sum_i \sum_t \frac{1}{\sigma_{it}^2} (y_{it}^* - a^* - c(h_{it} - h_{it_0})) + \frac{1}{\sigma_b^2} \mu_b}{\sum_i \sum_t (\frac{1}{\sigma_{it}^2} + \frac{1}{\sigma_b^2})}, \frac{1}{\sum_i \sum_t (\frac{1}{\sigma_{it}^2} + \frac{1}{\sigma_b^2})} \right)$$

- Sample  $c$  from the following normal distribution,

$$c \sim N^- \left( \frac{\sum_i \sum_t \frac{1}{\sigma_{it}^2} (y_{it}^* - a^* - br_t) + \frac{1}{\sigma_c^2} \mu_c}{\sum_i \sum_t (\frac{1}{\sigma_{it}^2} + \frac{1}{\sigma_c^2})}, \frac{1}{\sum_i \sum_t (\frac{1}{\sigma_{it}^2} + \frac{1}{\sigma_c^2})} \right)$$

- $\pi(a|\cdot) = \Gamma(2, 1 - \sum_{i=1}^i \log \delta_i)$  and  $\pi(b|\delta_i) = \Gamma(2, 1 + \sum_{i=1}^i \delta_i)$ .
- $\pi(\gamma_i|\cdot) = \Gamma(\frac{iT+1}{2}, \frac{1}{2}(\sum_{i=1}^i \sum_{t=1}^T \delta_i (y_{it}^* - a^* - br_t - c(h_{it} - h_{it_0})))^2)$ ,  $\sigma_{it}^{-2} = \delta_i \gamma_t$

### 3.2 Linear Regression Method

Linearising (5) and (7), allows us to estimate the coefficients using MLE. Let  $y_{mn}$  be a prepayment or a default rate in a certain region, where  $m$  represents the time of the origination of a loan and  $n$  represents the the duration till either prepayment or default. Thus the model for each region (there are 9 regions in this study) could be written in the following way:

$$y_{m,n}^P - \log(\min(n, 3)) = \log(a) + \beta(r_{m+n} - r_m) + \gamma(h_{m+n} - h_m)/h_m, \quad m = 1, 2, \dots, 19, \quad n = 1, 2, \dots, 10 \quad (8)$$

$$y_{m,n}^D - \log(\min(n, 3)) = \log(a) + \gamma(h_{m+n} - h_m)/h_m, \quad m = 1, 2, \dots, 19, \quad n = 1, 2, \dots, 10. \quad (9)$$

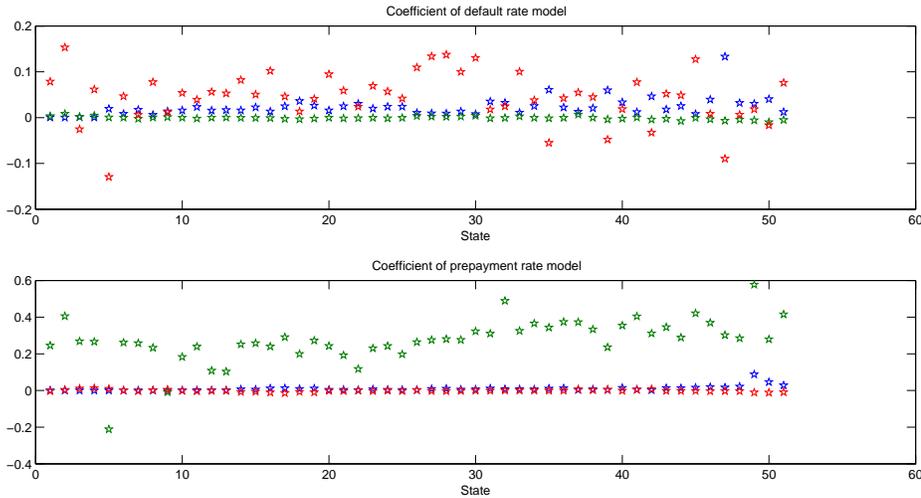


Figure 6: Estimated coefficients for prepayment and default rates for 51 state

Regions	1 ENC	2 ESC	3 MA	4 MT	5 NE	6 PAC	7 SA	8 WNC	9 WSC
HPI trend	1.9	1.6	2.3	1.5	2.8	2.2	1.6	1.3	1

Table 1: HPI Trends

We fit the above model for the historical prepayment and default rate data from all 9 regions separately and estimate the unknown parameters  $\log a$ ,  $\beta$ ,  $\gamma$ . For the whole data set, namely 51 state, estimated coefficients are illustrated in Figure 6.

For the estimating the coefficients in the prepayment rate and default rate model, we use standard linear regression with the yearly house price index, 30-year-loan-interest rate and the relative the prepayment rate (default rate), where the relative rates is a  $19 \times 10$  matrix in which the  $(i,j)$ -entry give us the probability of the borrowers of the loan initiated at year  $1980+i$  prepaying (default) at year  $1980+i+j$ .

Because of the limiting of the data set, we can only suppose for each regions, only one model for the prepayment (default) rate exist (i.e. the coefficients do not depend on time or the age of the loan).

After linear regression, we can see that the coefficients tell us if the interest rate get higher, borrowers intend to refinance their loan (i.e. the prepayment rate increase) which is coincide with the empirical results for the prepayment of the loan.

## 4 Modeling House Price Bubbles

There are two components to the HPI time series: linear trend and long term housing bubble. The linear trend is well modeled using the SDEs described above. We extract the linear trend,  $\theta_i$  in region  $i$ , from historic data.

The linear trend explains the growth and variation of the housing index during "normal" times, but it can not account for extremel events such as the rise and burst of a housing bubble. We create a second stochastic process to specifically model the housing bubble. For simplicity, a bubble is assumed to have the shape of a normal distribution. A bubble has three parameters: time gap between adjacent bubbles, the amplitude of a bubble, and the duration of a bubble. Each of these is a random variable. We extract their mean values from historical data: the average gap between bubbles is 10 years; at its peak, a bubble increases the housing index by 150 points; a bubble lasts on average 10 years from start to finish.

We use historic data to extract the inter-connections between bubbles in different regions. We observe that the amplitude of the bubble tends to be the largest in Middle Atlantic, New England, and Pacific regions. Moreover, the smaller bubbles in other regions tend to lag behind these regions which drives the economy

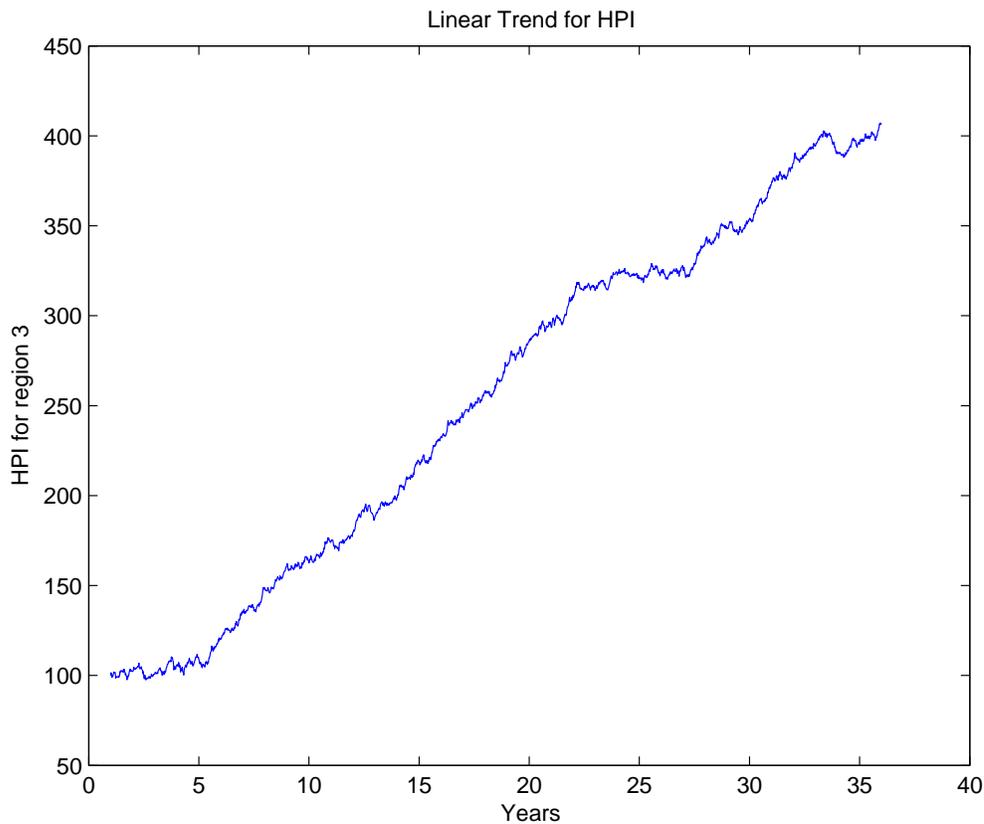


Figure 7: Linear HPI trend.

and housing market for the entire country. These information are captured in the distribution of the random variables corresponding to the starting time, amplitude and period of the bubbles.

Overview of the simulation platform

1. Use historic data to obtain the relations between HDI and interest rate (input) and prepayment and default rate (output).
2. Use historic data to capture the characteristics of HDI and interest rate. The historic information is embedded in SDE with oscillation for the housing index and in SDE with jump for the interest rate.
3. Once we have the stochastic models of HDI and interest rate, use these to generate a large number of samples of future HDI and interest rate. Each sample represents a future scenerio.
4. For each scenerio, use the simulated HDI and interest rate to generate the predicted prepayment and default rate.
5. Use the prepayment and default rate to calculate the expected revenue due to premium and loss due to default for each housing contract.
6. Aggregate over all housing contracts to calculate the total income stream.
7. Iterate steps 4-6 over all scenerios to generate a distribution of future income.

As an example, consider a bucket of 1000 housing contracts, each contract is for a loan of 150,000 USD at 0.0065 premium rate and 15 percent Optional Guarantee. Figure ??? shows the income stream for this

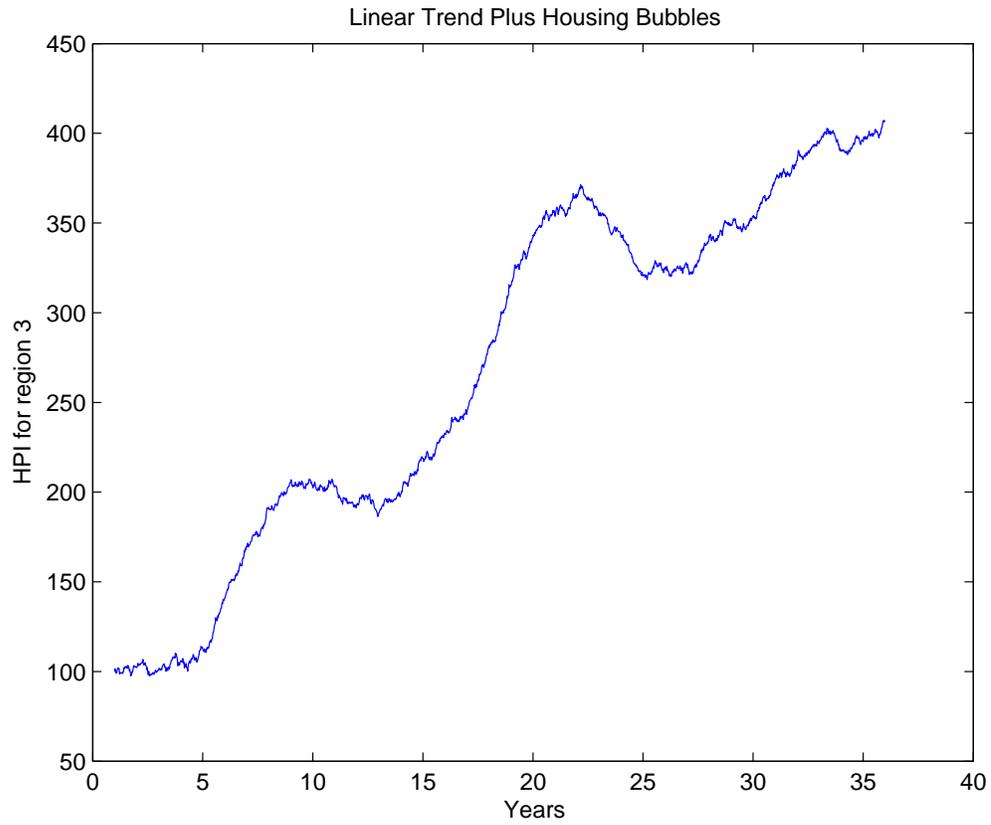


Figure 8: Linear HPI trend with bubbles.

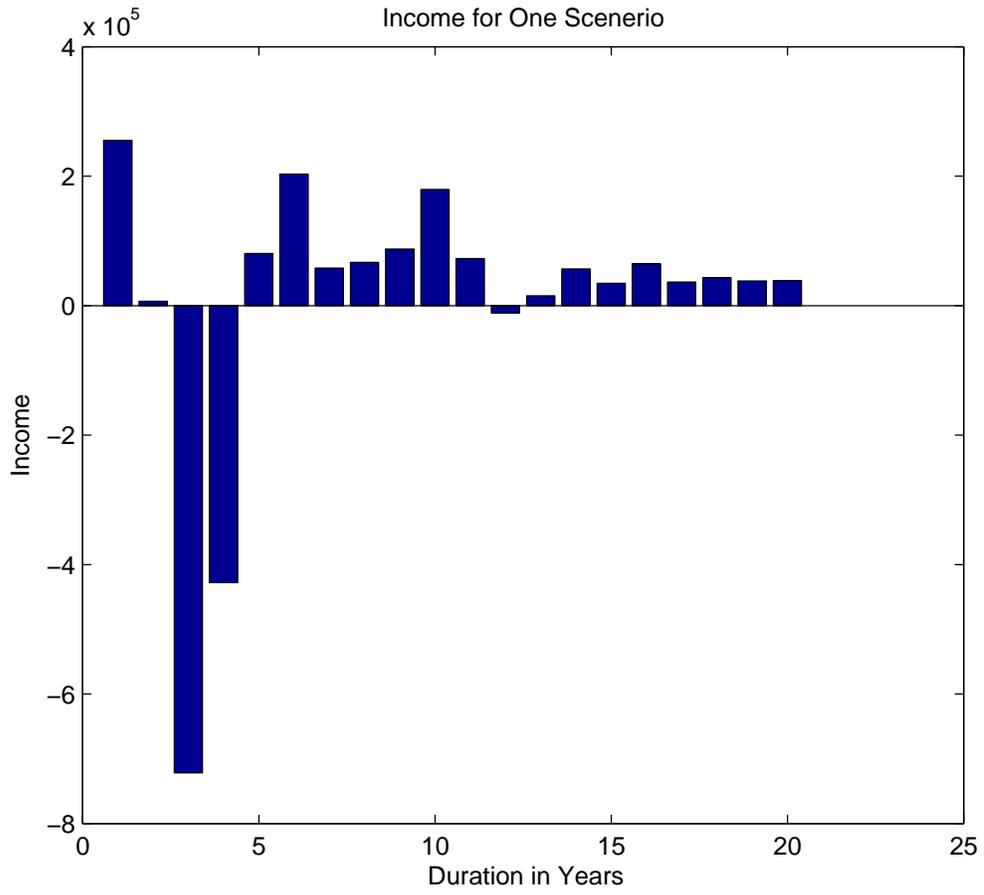


Figure 9: Income Stream for One Scenerio.

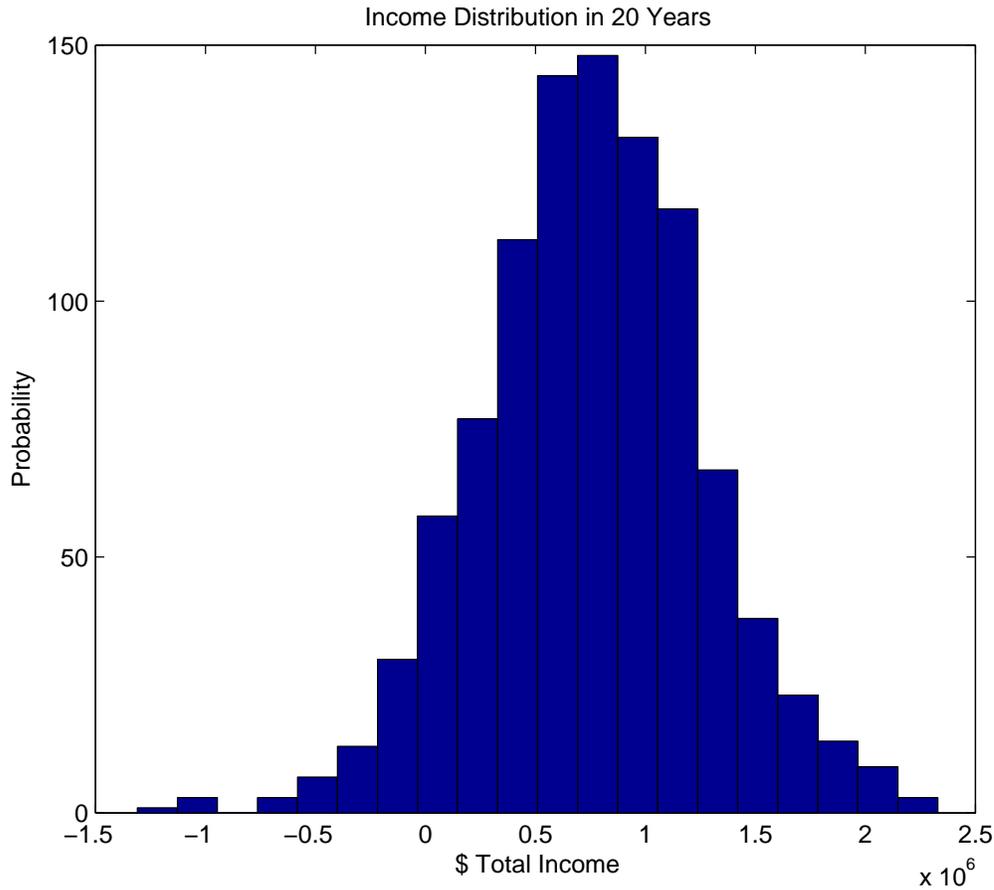


Figure 10: Income Distribution for Low Risk Model.

bucket over 20 years for one particular scenario (i.e. for one particular pair of simulated HPI and interest rates). We see that the net income is negative in some years due to large number of defaults.

We compute the distribution of total income during 20 years of this bucket over 1000 scenarios. We present two sample distributions: one low-risk estimate and one high-risk estimate. The low-risk model assumes that the probability of encountering a large bubble over the duration (20 years) is small. While the high risk model assumes that the probability of a large correlated bubble over all the regions is large. For the low risk model, the probability of the income being negative (i.e. firm losing money) is 0.067. For the high risk model, this probability is 0.35. In our simulator platform, we allow the user/analyst to select various risk levels and investigate the resulting income distribution. We anticipate that different firms have different level of "acceptable" risk and hence we make risk into a turnable dial.

## 5 Binary Time Series Analysis of House Price Indices.

**Order Estimation** Before performing order estimation, we map the HPI time series to a finite set of two states in the following way:

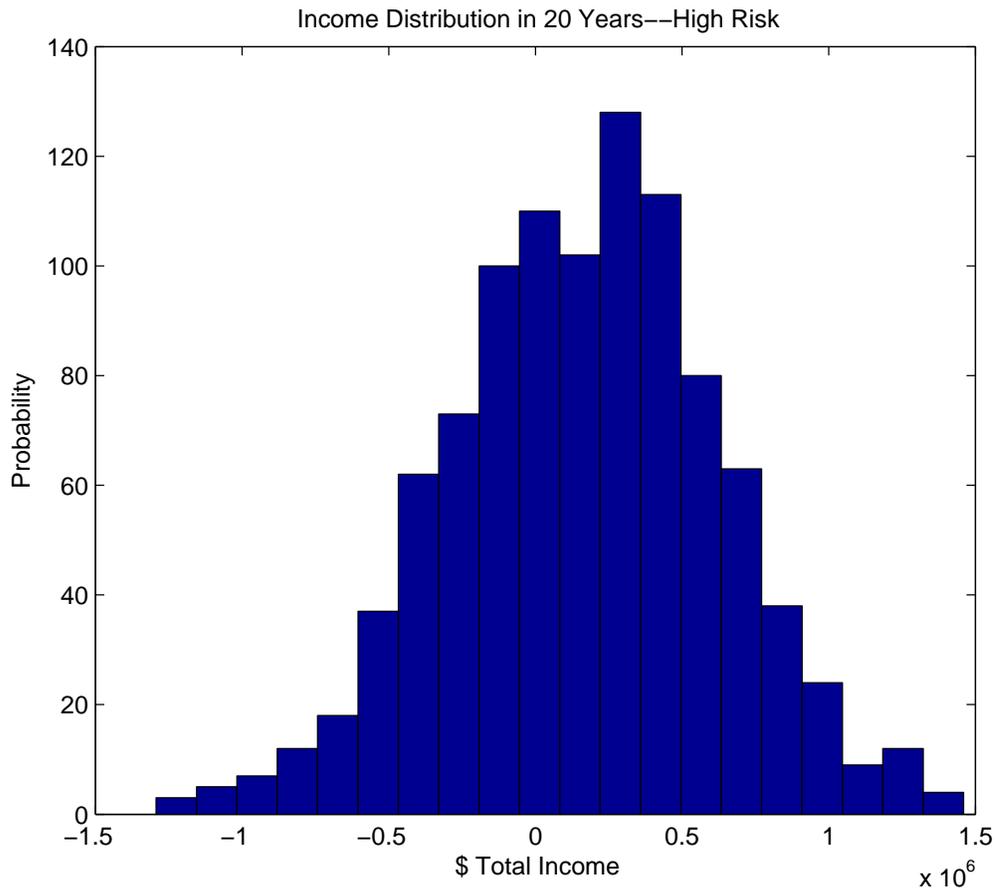


Figure 11: Income Distribution for High Risk Model.

Let  $x_1, x_2, \dots, x_n$  be the change in the HPI, i.e.  $x_i = HPI_i - HPI_{i-1}$ , we then define  $y_n$  as:

$$y_n = \begin{cases} u & x_n \geq 0 \\ d & x_n < 0 \end{cases}$$

where  $u$  and  $d$  stand for up and down movement of the house price index respectively.

After mapping data onto  $u$  and  $d$  states, we use the provably consistent BIC method [2] to estimate the order of the binary time series. We calculate

$$f(k, N) = L_k - (1/2)(q-1)q^k \log N \quad (10)$$

where  $L_k$  is the log likelihood of the mapped  $q(=2, \text{ in our case})$  state time series data computed using MLE estimates of the  $k$ -th order Markov Chain transition probabilities and  $N$  is the sample size. Looping over  $k$  from 0 to  $K$ , the BIC order estimate equals the value of  $k$  that maximizes (10).

Performing the above order estimation for nine different sets of HPI data based on nine different regions, we see that the up and down movements in HPI is best estimated by a first order Markov Chain for all the regions.

**Maximum Likelihood Estimation (MLE) for  $k$ -th order Markov Chain** Assume that each  $X_j$  is either  $u$  or  $d$ , so that the entire data set is an  $u/d$  sequence of length  $N$ . For a  $k$ -th order Markov chain,  $X_j$  is either  $u$  or  $d$  and this outcome depends on  $k$  outcomes prior to the  $j$ th outcome. Hence, this Markov property can kick in only if  $j > k$ . Let  $S_i$  be a sub-sequence of  $k$  outcomes prior to the  $j$ th outcome. Since each of these  $k$  outcomes in the sub-sequence can be  $u$  or  $d$ , there are  $2^k$  possible sub-sequences  $S_i$ . Let these sequences be denoted by  $\{S_i\}_{i=1}^{2^k}$ . Let us scan the given sequence  $\{X_j\}_{j=1}^N$  from left to right and record the following  $2^{k+1}$  numbers:

$$\begin{aligned} n_{S_i u} &= \# \text{ of times we observe "S}_i\text{u"} \\ n_{S_i d} &= \# \text{ of times we observe "S}_i\text{d"} \end{aligned}$$

Let  $p_{S_i u} = P(u|S_i)$ ,  $p_{S_i d} = P(d|S_i) = 1 - p_{S_i u}$ ,  $p_k = P(\{X_j\}_{j=1}^k)$ . Putting everything together, the likelihood function for an  $u/d$  sequence of length  $N$  is

$$L(p_{S_1 u}, p_{S_2 d}, \dots, p_{S_{2^k} u}) = p_k \prod_{k=1}^{2^k} (p_{S_i u})^{n_{S_i u}} (1 - p_{S_i u})^{n_{S_i d}}$$

Taking logarithms of both sides, we obtain the log likelihood  $L_k$  for all  $k$ :

$$\log L(p_{S_1 u}, p_{S_2 u}, \dots, p_{S_{2^k} u}) = \log p_k + \sum_{k=1}^{2^k} \log(p_{S_i u})^{n_{S_i u}} \log(1 - p_{S_i u})^{n_{S_i d}} \quad (11)$$

**First Order Estimates** The first order transitional probabilities as estimated by MLE method gives us the estimates for the probabilities  $P(u|u)$  and  $P(u|d)$ . We further estimate  $P(d|u)$  and  $P(d|d)$  as  $1 - P(u|u)$  and  $1 - P(u|d)$  respectively for each of the nine regions. It is observed that the regions with a pronounced bubble have significantly higher values of  $P(u|u)$  and  $P(d|d)$  as compared to the regions without a bubble.

**Order Estimation for more States( $>2$ )** We can further convert the HPI time series into a discrete time series with more symbols depending on percentage upward and downward movements and estimate the order of the new Markov Chain with (10) with  $q$  being the number of states chosen. We can then find the transition probabilities to predict the percentage upward and downward change (magnitude of movement) rather than just the upward and downward movement.

## 6 Conclusions

We have analyzed the historic data and extracted relations between housing index and interest rate versus default and prepayment probabilities. We have created projection tools based on these historic results. We have created a technology platform that allows the users to select a risk profile and explore his future income range. Historic income and default data agrees well with our predictions.

## References

- [1] J. C. Hull, *Options, Futures, and Other Derivatives*, Prentice-Hall, Upper Saddle River, 2000.
- [2] I. Csiszár and P. C. Shields, *The consistency of the BIC Markov order estimator*, *Annals of Statistics*, 28 (2000), pp. 1601–1619.

# Resource Issues Impacting National Security

Aaron Brown<sup>1</sup>, Chad Griep<sup>2</sup>, Magathi Jayaram<sup>3</sup>, Karthik Jayaraman Raghuram<sup>4</sup>, Kirk M. Soodhalter<sup>5</sup>, Duy Vu<sup>6</sup>, Jelani Wiltshire<sup>7</sup>

Problem Presenter:

John Peach

MIT Lincoln Laboratory

Faculty Mentor:

Mansoor Haider

North Carolina State University

## Abstract

## 1 Introduction and Motivation

Resource depletion can influence the policy decisions of nations, and this is especially true now that many natural resources are being severely depleted. Oil has passed peak production which poses a threat to the transportation costs associated with the import and distribution of food and other goods in a nation. This, coupled with the impending rise of large middle class populations in developing nations, will increase pressure on these resources as demand for personal vehicles increases. Burning of fossil fuels releases carbon dioxide further exacerbating the problem of global climate change. This will have further effect on the food supply placing further pressure on vulnerable states, leading to the creation of refugee populations which can feed extremist movements. Physical location and control of essential resources by competing or hostile nations also exerts influence on policy decisions. Current national discussion has centered on the reduction of dependence on the use of fossil fuels through technological innovation, renewable energy sources, and improvement in efficiency of energy use. However, there are concerns that the resources needed to build these new technologies may be scarce.

The interactions of these issues highlights the need for a model relating the interdependencies of population, resource depletion, food production, renewable technologies, and energy generation. Starting from the preexisting implementation of the World3-03 Model as described *The Limits to Growth* [6]. Based on experiments with this initial model and some insights gained from preliminary research, we sought to reduce this model to its essential relationships. This turned out to be rather complicated and unfeasible in our short timeframe. Instead, we began the process of sensitivity analysis to determine the sensitivity of certain outputs to perturbations of certain inputs, choosing which inputs and outputs to test based on policy questions about which we wished to probe the model. We answer some of our questions and chart a course for future study of this model in a standardized fashion.

## 2 Projecting Future Resource Supply

We know that all mineral resources are in finite supply on earth, but it is difficult to project what quantity of these resources is yet to be discovered. This is of primary importance for fossil fuel resources upon which much of our economic activity and food production rests. The most obvious example is the need to project how much oil is yet to be extracted from the earth. However, even if the world shifts the burden of energy

---

<sup>1</sup>Tufts University

<sup>2</sup>University of Rhode Island

<sup>3</sup>Utah State University

<sup>4</sup>University of California - Santa Barbara

<sup>5</sup>Temple University

<sup>6</sup>Penn State University

<sup>7</sup>Florida State University

generation to renewable technologies, we must ask if the resources needed to build these new technologies. For example, solar panels and electric car batteries use resources that may not be abundant enough to sustain a large scale change in our energy production. This issue may be further exacerbated by the fact that some of the resources currently used in the fabrication of these technologies are also involved in the production of other technologies. For example, silicon, which is used to produce solar cells, is dominated by the semiconductor industry and as a raw material for producing silicones. Indium and germanium, also used to produce solar cells, are used in the construction of LCD flat screen panels and televisions.

We use the Hubbert Linearization in order to project the total cumulative quantity of a resource that will be extracted. If we assume that the amount of a resource extracted follows a logistic curve with respect to time, the Hubbert Linearization exploits the linear nature of the logistic model to represent yearly production as a percentage of total production as a linear function of total cumulative production. In other words, let  $Q(t)$  be the function representing total cumulative amount of a resource extracted at a particular time where the variable  $t$  is time in years. Then the assumption that  $Q$  grows logistically is represented by the differential equation

$$P(t) := \frac{dQ}{dt} = KQ \left( 1 - \frac{Q}{F} \right)$$

where  $K$  is the logistic growth rate and  $F$  is the total amount of resource to be possibly recovered. Dividing both sides by  $Q$  gives

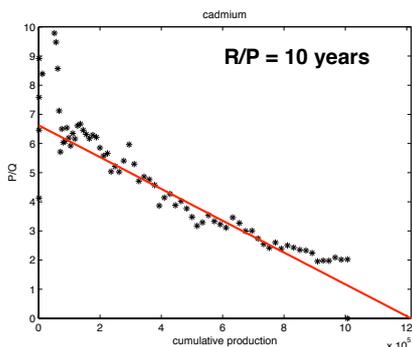
$$\frac{P}{Q} = K \left( 1 - \frac{Q}{F} \right)$$

which represents percentage of yearly production as a linear function of cumulative production at time  $t$ . This allows us to project this percentage into the future and to estimate the total available resource that can possibly be extracted. From this along with data on yearly consumption of the resource, we can estimate how many years supply we have of the resource. This idea was originally suggested by Hubbert in 1982 to study mineral resource supplies and applied more recently to the extraction of oil.

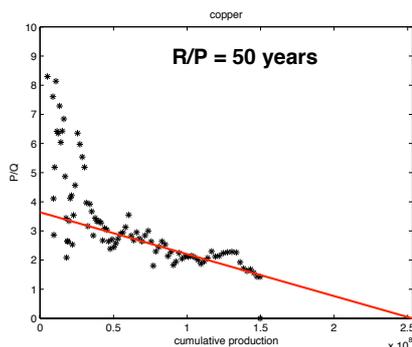
The actual computation and construction of the linearization is straightforward. The actual data is freely available on the United States Geological Survey's (USGS) website. We import the data for each resource into Matlab. We computed the cumulative resource production for each year. We can then plot cumulative resource production per year against yearly production as a percentage of cumulative production with cumulative production as the independent variable. We then observed that it is not prudent to simply fit a linear model to these points. In the first year of production, the amount of a resource produced is 100% of total production; and in the first few years, this percentage drops at a steep rate before stabilizing. To get a better idea of the linear relationship between these two quantities, we discarded the data points corresponding to the 25% of the interval  $[0, Q_{max}]$  where  $Q_{max}$  is the cumulative resource production for the last year of recorded data. The actual model was constructed using Matlab's `polyfit()` function where we specified that we want a degree 1 polynomial fit to the data. Where this line intersects the horizontal axis corresponds to the total amount of the resource projected to be found before exhausting underground supplies. Then by computing the total amount of resource yet to be extracted and dividing by the amount produced in the last recorded year, we estimated how many more years we can expect to produce more of the resource. The most worrying projects are for phosphate and germanium. Besides being an element used in the production of renewable energy, phosphate is also extremely important in the production of fertilizers in agriculture. Germanium demand is increasing both because of its use in renewable energy technologies and in flat screen televisions and panels.

The veracity of our Hubbert Linearization rests on the assumptions that the supply of each resource is, in fact, governed by a logistical model and that the data we acquired from the USGS website is complete. We could not test the second assumption, but we did investigate the first by testing how well the linearization fits the existing data. Therefore, we computed the  $\mathcal{R}^2$ , which is the proportion of variance explained by the linear fitting model. An  $\mathcal{R}^2$  closer to 1 implies better fitness of the model. We also computed the  $p$ -value to test the significance of the predictor cumulative production, where  $p < .05$  is taken to be significant. This analysis shows that the predictions by Hubbert Linearization for lithium and copper are not good fits for the data. This may imply that our hypothesis about copper and lithium production growing logistically may be incorrect. We may undertake further testing to verify models which are not invalidated by  $\mathcal{R}^2$  by *back-casting*, which is to construct a model using historical data and see if it accurately predicts (approximately) data from the years following the historical data.

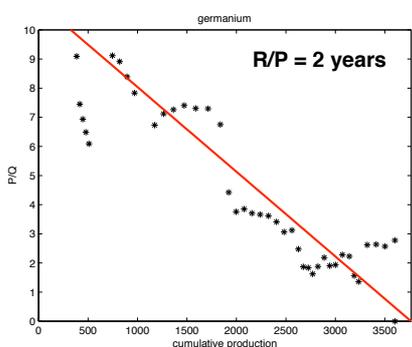
Figure 1: Hubbert Linearization for Various Resources.  $R/P$  is projected remaining resources divided by production in the last year of recorded data which gives an idea of amount of productive years left.



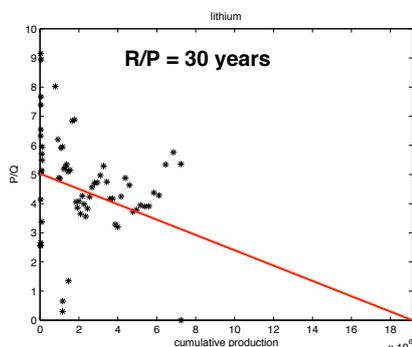
(a) Hubbert linearization projects cadmium to last another 10 years.  $\mathcal{R}^2 = .9105$  and  $p < 1.9 \times 10^{-16}$



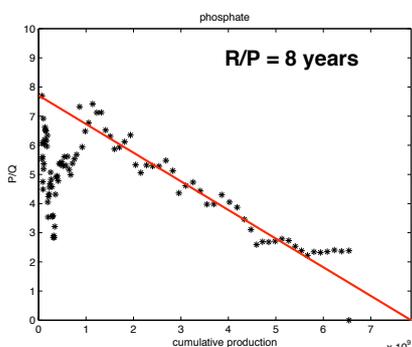
(b) Hubbert linearization projects copper to last another 50 years.  $\mathcal{R}^2 = .4837$  and  $p < 2.2 \times 10^{-9}$



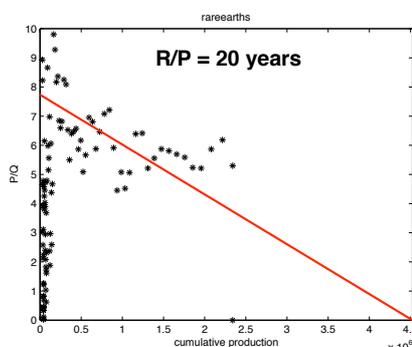
(c) Hubbert linearization projects germanium to last another 2 years.  $\mathcal{R}^2 = .7682$  and  $p < 6.7 \times 10^{-12}$



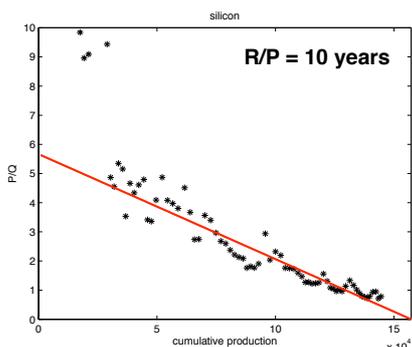
(d) Hubbert linearization is a poor fit to the date with  $\mathcal{R}^2 = .1135$  and  $p = .03347$



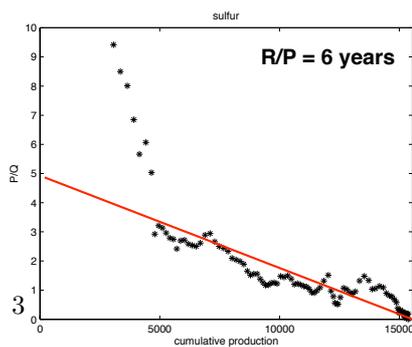
(e) Hubbert linearization projects phosphate to last another 8 years.  $\mathcal{R}^2 = .9289$  and  $p < 2.2 \times 10^{-16}$



(f) Hubbert linearization projects various rare earth elements taken as a whole is a poor fit with  $\mathcal{R}^2 = .1023$  and  $p = .07656$



(g) Hubbert linearization projects silicon to last another 10 years.  $\mathcal{R}^2 = .8923$  and  $p < 2.2 \times 10^{-16}$



(h) Hubbert linearization projects SULFUR to last another 6 years.  $\mathcal{R}^2 = .8923$  and  $p$ -value  $< 2.2 \times 10^{-16}$

### 3 Fossil Fuels and Carbon-Dioxide Emissions

Increases carbon dioxide ( $CO_2$ ) levels in the atmosphere have been linked with global warming. One of the ways that  $CO_2$  enters the atmosphere is through the use of fossil fuels. We are interested in modeling the effects of fossil fuel consumption on the trend of annual  $CO_2$  emissions. We propose study this by using a univariate  $ARIMA(p, d, q)$  model for  $CO_2$  emissions and a regression-time series model for  $CO_2$  emissions on fossil fuel consumption. We will use annual (1980-2006)  $CO_2$  emissions and fossil fuel consumption data for coal and natural gas obtained from the Energy Information Administration website [3]. If  $X_t$  follows an  $ARIMA(p, d, q)$  then it is defined as follows [9]

$$\phi(B) (1 - B)^d X_t = C + \theta(B)a_t$$

where

$$\begin{aligned}\phi(B) &= 1 - \phi_1 B - \dots - \phi_p B^p \\ \theta(B) &= 1 - \theta_1 B - \dots - \theta_q B^q\end{aligned}$$

and  $d$  is the trend differencing order,  $\{a_t\}$  is a zero mean white noise process,  $p$  is the Autoregressive (AR) order, and  $q$  is the Moving-average (MA) order.

The autocorrelation,  $\rho(k)$ , in the residuals,  $\{a_t\}$ , will be tested using the Ljung-Box Q-test: [9]

$$\begin{aligned}H_0 : \rho(1) = \rho(2) = \dots = \rho(K) = 0, \\ H_1 : \text{At least one } \rho(k) \neq 0\end{aligned}$$

$$Q = n(n+2) \sum_{k=1}^K (n-k)^{-1} \hat{\rho}(k)^2 \approx \chi_{K-p_1}^2$$

where  $p_1$  is the number of significant parameters(not including the intercept) and for a random sample of size  $n$ ,  $\{y_1, y_2, \dots, y_n\}$ :

$$\hat{\rho}(k) = \frac{\sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2}$$

for  $k = 1, \dots, \frac{n}{4}$ .

The regression-times series model used here has the following form:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \dots + \beta_k X_{kt} + N_t, t = 1, \dots, sn$$

where  $\beta_1, \dots, \beta_k$  are estimated using Least Squares regression, the random error  $\{N_t\}$  is further modeled by an  $ARIMA(p, d, q)$  process for  $d=0$ .

The first model that was estimated using a univariate  $ARIMA(1,0,2)$  for Worldwide  $CO_2$  emissions from coal consumption (Million Metric Tons),  $X_t$ , from 1980-2006:

$$(1 - 0.99842B) X_t = (1 + 0.89937B + 0.3851B^2) a_t \quad (1)$$

The parameter estimates of the MA part and the AR part of the model were significant at the 0.05 level. The mean was not significant, so it was dropped from the model. The Ljung-Box Q-test rejects the null hypothesis of no autocorrelation in the residuals at lag 6 and lag 24 at the 0.05 level of significance. This violates the assumption of no autocorrelation in the residuals and suggest that this model might not be consistent or adequate in explaining the variation in the data. The AIC is 1127.661.

The second model that we estimated was a regression-time series model of Worldwide  $CO_2$  emissions,  $\{Y_t\}$ , on Coal consumption (short tons),  $\{X_t\}$  from 1980-2006:

$$Y_t = 1.72222X_t \quad (2)$$

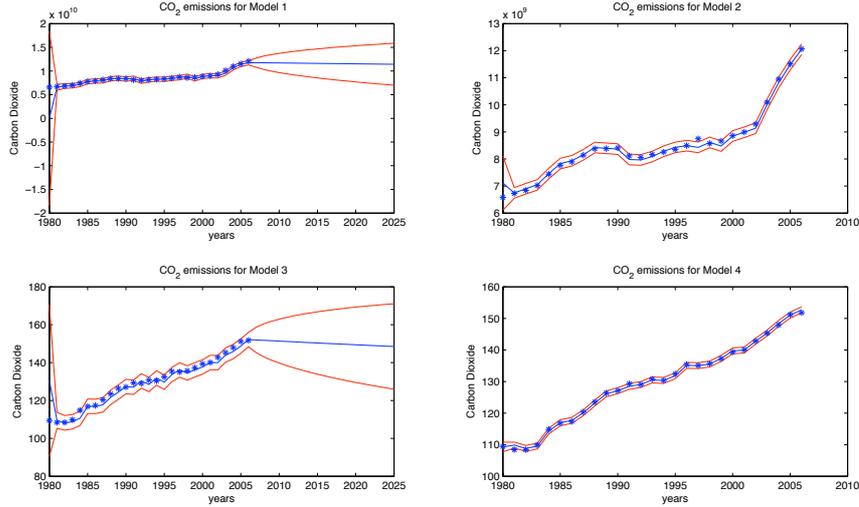


Figure 2:  $CO_2$  forecasts in model 1 through model 4. with a 95% confidence interval.

and

$$N_t = 0.98067N_{t-1} + a_t \quad (3)$$

The estimates for this model were significant at the 0.05 level and the AIC was 1076.481. The Ljung-Box Q-test rejects the null hypothesis at lag 6, but the remaining lags are not significant.

The third model we estimated was a univariate ARIMA(1,0,1) for  $CO_2$  emissions from Natural gas consumption (Million Metric Tons),  $\{X_t\}$ , from 1980-2006. Before estimating the model we used a Box-Cox transformation of 0.5 to stabilize the variance. Then we estimated the following model:

$$(1 - 0.9903B)X_t = 1.26632 + (1 + 0.46783B)a_t \quad (4)$$

The AIC for this model is 120.2246. The Ljung-Box Q-test rejects the null hypothesis for all the lags, suggesting strong autocorrelation in the residuals.

The fourth model that we estimated was regression-time series model for  $CO_2$  emissions,  $Y_t$ , on Natural gas consumption (Billion Cubic Feet),  $X_t$  from 1980-2006. Using transformed  $CO_2$  emissions from the third model we estimated the following model:

$$Y_t = 64.40581 + 0.000847X_t \quad (5)$$

and

$$(1 - 0.79171B)N_t = a_t \quad (6)$$

The AIC of this model is 40.344. The Ljung-Box test failed to reject the null hypothesis for all lags. This suggests that there is no autocorrelation among the residuals and we have a white noise process.

Several tests of Normality: Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling showed that residuals from all 4 models depart from normality at the 0.05 level. The maximum likelihood estimation method was used to estimate the parameters of the models. The parameter  $d$  was taken to be zero in all four models. The regression-times series appears to fit the data better than the univariate time series model in both the Natural gas data and the Coal data because the regression time series models have lower AIC values 40.344 and 1076.481, respectively.

There are several problems with these model. One of the major problems is the size of the data set. The data set contains only 27 data points so the model so the problems that we can run into in making forecasts include potentially overfitting our the data and not being able to make accurate predictions outside of our data range. The figure also shows that the variance estimate appears to grow with the number of steps forecasted into the future. This would imply that the forecasts could be unstable.

## 4 The World3-03 Model

### 4.1 Review of the World3-03 Model

The World3-03 Model was developed by Meadows et al. in 1972 to model the interactions of Earth’s human populations with respect to resource consumption, agriculture, economic activity, and other factors to study the interactions between the Earth’s natural resources and human activities. Their queries of this model formed the basis of the book *The Limits to Growth* and the update of the model was used in the thirty-year update *Limits to Growth: The 30-Year Update* [6]. This model is implemented in the Vensim modeling software. Vensim allows the user to define relations between different inputs and outputs through GUI interface manipulating a flow chart-like diagram. The World3-03 Model Vensim implementation is contained in fourteen interconnected pages organized by node categories, e.g. demographics, energy demand, food production, land use, carbon emissions; it is quite large and highly interdependent. Therefore, probing the model to determine important nodes and loops with the goal of model reduction is much too ambitious in the context of a ten-day workshop. The purpose of the world model is not to make specific predictions, but to explore how exponential growth interacts with finite resources. The main systems comprising the model are

- the food system,
- the industrial system,
- the population system,
- the non-renewable resources system,
- the pollution system.

The sheer vastness of the model, hinders ones ability to gain a complete understanding of the interrelations between the over 400 variables. See Figure 3. Essentially, one is relegated to black box analysis, where one enters input and receives output without any knowledge of the inner workings of the model.

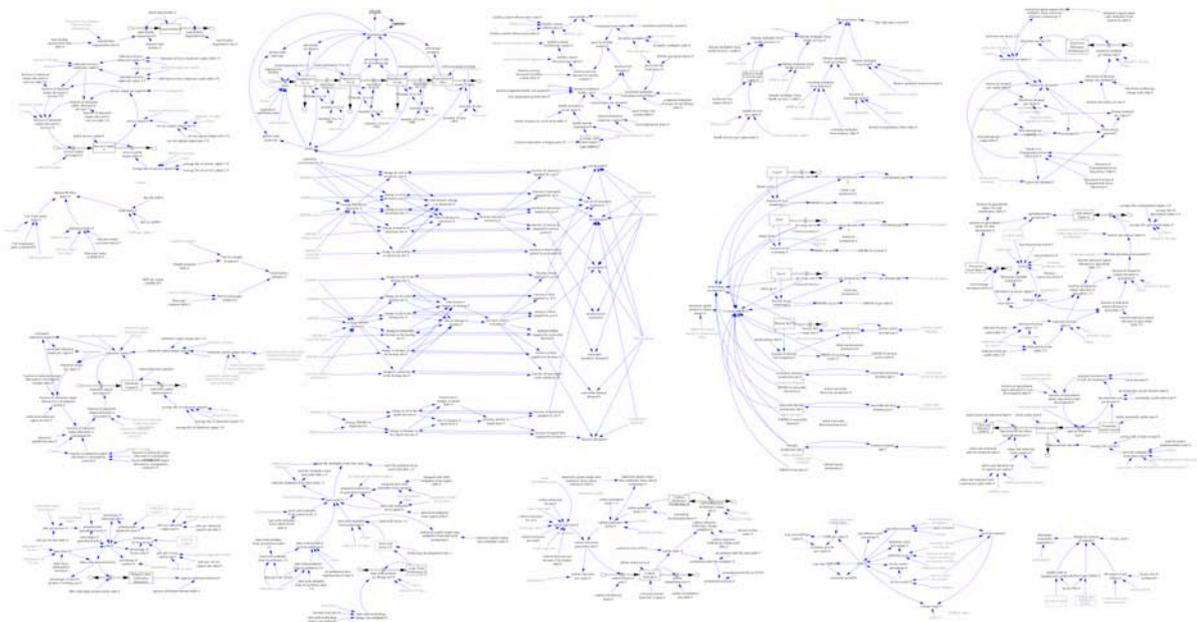


Figure 3: The Full World3-03 Model as Implemented in Vensim.

In order to facilitate gaining a rudimentary understanding of the relational dependencies in the model, we develop a “simplified” structural version of the model. Redundant nomenclature and overly detailed

components have been removed, while the general structure of the model has been preserved. Although reducing the model highlights the causal relationships, the obvious deficiency is that the functional relationships become less clear. See Figures 4, 5, and 6.

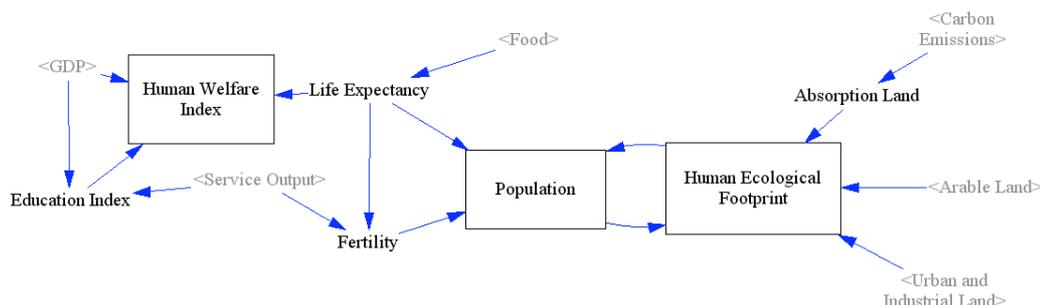


Figure 4: Demographics Diagram in the Reduced Model

From this reduced version of the World3-03 Model, we are able to pinpoint possible areas of improvement. For example:

- Total Fuel for Transportation is not a function of Population and is primarily dependent upon Service Output and Industrial Output. Since a large amount of transport fuel is used for personal transportation, we should include a Private Use Factor for the Total Fuel for Transportation quantity.
- Total Fuel for Transportation and thus Liquid Fuel Demand is not a function of the Gross Domestic Product, while Heat Demand and Electricity Demand are. We can assume that transportation and thus the consumption of liquid fuel would increase as GDP per capita increases.
- Total Fuel for Transportation only affects Liquid Fuel Demand and not Electricity Demand. Thus, the model does not incorporate the possibility of electric powered transportation. This motivated some modifications we made to the model which will be discussed in the next section.
- Carbon Emission is a function of Oil Usage, but the model does not take into account specific methods of oil combustion. Perhaps, Oil Usage can be subdivided into Oil Usage for Transportation, Oil Usage for Heat, and Oil Usage for Electricity.
- The current model does not factor in the electricity conversion loss which occurs during the generation, transmission, and distribution of electricity.

Donella Meadows, one of creators of the World3-03 Model, acknowledges that certain aspects of the model may not be ideal. In [5], p. 129, Meadows states:

We have a mixed degree of confidence in the numerical parameters of the model; some are well-known physical or biological constants that are unlikely to change, some are statistically derived social indices quite likely to change, and some are pure guesses that are perhaps only of the right order of magnitude. The structural assumptions in World3-03 Model that I consider most dubious and also sensitive enough to be of concern are:

- the constant capital-output ratio (which assumes no diminishing returns to capital),
- the residual nature of the investment function,
- the generally ineffective labour contribution to output.

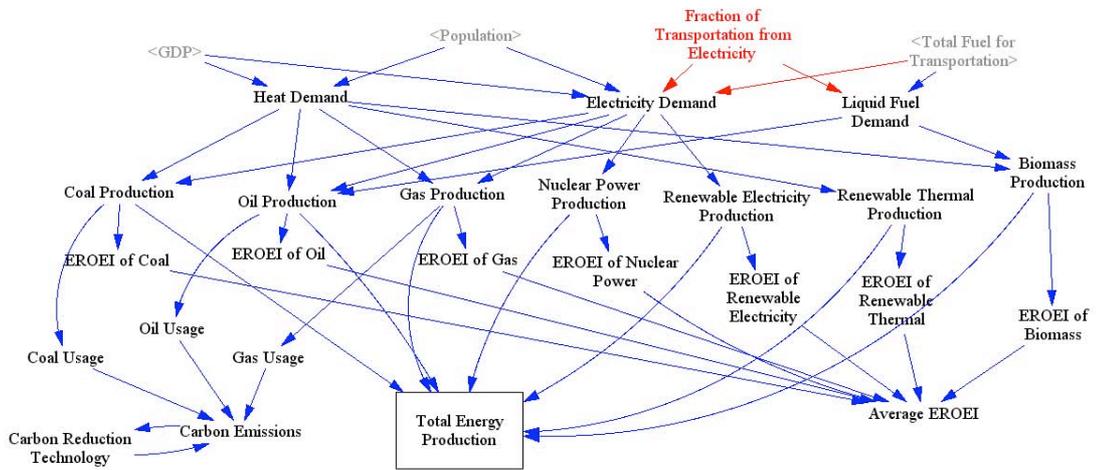


Figure 5: Energy Diagram in the Reduced Model

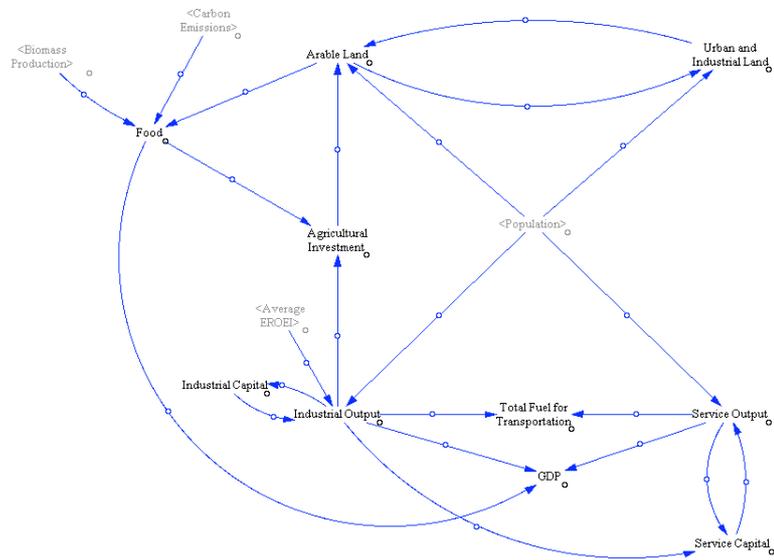


Figure 6: Industry, Agriculture, and Services Diagram in the Reduced Model

## 4.2 Our Additions to the Model

To investigate the influence of new technologies in the existing model, we factored the increased use of electric vehicles into the model. In the original model, industry and agriculture create a demand for liquid fuel, which is met by oil production and a small contribution by biomass (presumable ethanol based technologies). To alleviate the demand on oil we forced a fraction of transportation to be powered via electricity rather than liquid fuels. This would correspond to electric car technology and electric mass transit. To make our model historically realistic and allow for development and implementation of new technologies, the fraction of transportation powered by electricity was gradually increased between the years 2000 and 2060. Ultimately we optimistically allowed for 30% of transportation to be powered via electricity. See Figure 4.2.

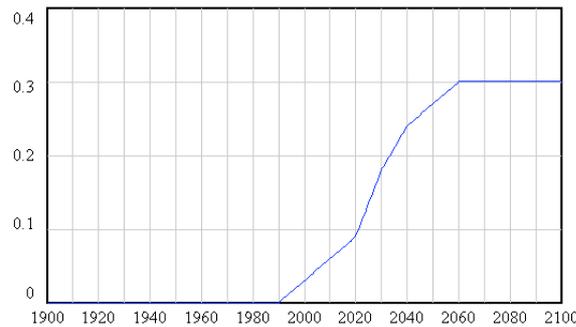


Figure 7: Fraction of Cars Powered by Electricity in Modified Model

In the Vensim model this is accomplished by the following modification to the “Energy Demand” page. This is demonstrated in Figure 4.2 with new nodes are highlighted in black.

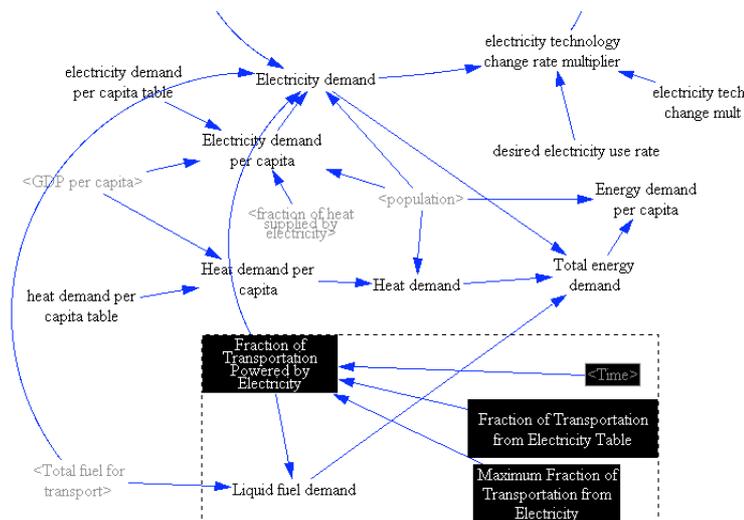


Figure 8: Modification to “Energy Demand” Page of World3-03 -2003 Model

The main result of adding electric cars into the model was to increase the carbon-dioxide levels in the atmosphere in the late 21st century. The reason for this is that the shift to electric cars creates a higher demand for coal in the latter half of the 21st century, and in the model, coal produces more CO<sub>2</sub> per unit of energy than oil (see Figure 4.2)

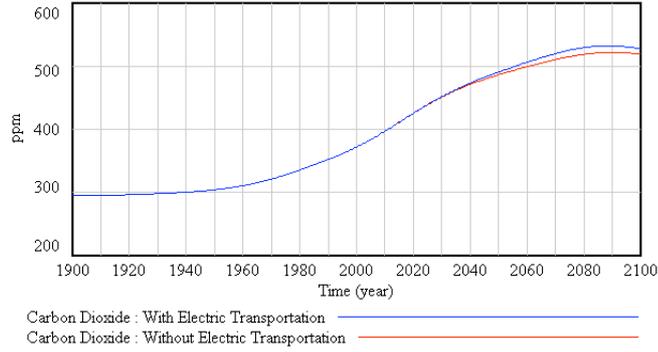


Figure 9: Carbon-dioxide levels projection in World3-03 Model

## 5 Sensitivity Analysis

### 5.1 Review of Sensitivity Analysis

Since the size of the Vensim model prohibits direct study in the relationships represented therein, we must use a method of probing the model as a black box for which we do not know the internal function. We employ a method known as sensitivity analysis. Essentially, the Vensim implementation of the World3-03 Model represents a black box with known inputs and outputs. Sensitivity analysis refers to the study of the response of the output of a dynamical system with respect to changes in the input parameters. Typical methods for sensitivity analysis include:

1. Local Methods, such as those involving the study of derivatives
2. Sampling Methods, including Monte Carlo filtering
3. Variance based methods
4. Emulator based methods

Sensitivity analysis determines how sensitive a model is to the changes in the value of the parameters in the model and to changes in the structure of the model. Parameter sensitivity is usually performed as a series of tests in which different parameter values are set to see how a change in the parameter causes a change in the dynamic behavior of the system. By showing how the model behavior responds to changes in parameter values, sensitivity analysis is a useful tool in model building as well as in model evaluation.

Sensitivity analysis helps to build confidence in the model by studying the various uncertainties involved with varying the parameters. Sensitivity analysis can also indicate which parameter values are reasonable to use in the model. Experimenting with a wide variety of values gives us insight into the behavior of the system under extreme conditions. Discovering the system behavior greatly changes for a particular parameter helps us identify the leverage point in the model- a parameter whose value can significantly affect the behavior of the system.

A simple local method is where we study the terms of the Jacobian Matrix,

$$J_{i,j} = \left| \frac{\partial F_i}{\partial P_j} \right| \quad (7)$$

We use a center difference to approximate the partial derivatives, and use a local normalization. Normalization is important because comparison of entries in the Jacobian is meaningful only if all the derivatives are dimensionless.

We then have the sensitivity of an output  $F_i$  to the parameter  $P_j$  is given by:

$$S_{i,j} = \frac{\partial F_i}{\partial P_j} \times \frac{P_j}{F_i} \quad (8)$$

$$\approx \frac{F_i^+ - F_i^-}{P_j^+ - P_j^-} \times \frac{P_j}{F_i}, \quad (9)$$

where  $P_j^\pm = P_j \pm \delta P_j$  and  $F_i^\pm = F_i(P_j^\pm)$ .

Typically, derivative based sensitivity analysis is performed through automatic differentiation. However, in the case where we only have access to a black box, we sample the model at different values of interest, and approximate the derivative.

A more comprehensive method would involve Monte Carlo filtering, where we choose the probability density function for the parameter of interest. A random number generator generates samples of the parameter and the sensitivities are computed. This approach could not be used with the World3-03 Model since the trial version of Vensim does not permit automated parameter sweeps required for Monte Carlo simulations.

## 5.2 Consequences and Conclusions of Sensitivity Analysis

Because of the limitations of the software, it was unfeasible to run a large sensitivity analysis. However we were able to manually calculate relative sensitivities for a small set of variables and parameters in the model. Table 1 summarizes our results by showing the maximum (in absolute value) relative frequency for each variable-parameter pair over the time interval.

	Industrial Output	Life Expectancy	CO <sub>2</sub> Levels	Human Welfare Index
EROEI of Renewable Electricity (15–25)	0.8151	0.0655	-0.0063	0.2570
Fraction of transportation from electricity (25%–35%)	-0.1804	-0.0100	0.0178	-0.0876
Carbon reduction technologies (-.5–1.5)	-0.1489	0.0000	-0.0876	-0.0649

Table 1: Maximum relative sensitivities of a small set of variables and parameters in the model.

We may draw some policy conclusions from this table. First, we clearly see that increasing the EROEI of renewable electricity (presumably through new technology) has a strong (positive) influence on the industrial output and human welfare index. We also note that Carbon Reduction Technologies (in the range considered) have the strongest effect on reducing CO<sub>2</sub> levels. However examining the sensitivity dependence over time in Figure 10 reveals that the increase in EROEI and the proposed switch to electric powered transportation has a stronger influence in the early 20th century. This is consistent with the fact that carbon reduction technology is not set to take effect until quite late in the 20th century in the model.

Figure 11 compares the sensitivity of the human welfare index against the 3 parameters. We note that carbon reduction technologies has a negative impact on the human welfare index as it is assumed in the model that the development and production of new technologies diverts industrial capital away from GDP.

## 6 Search for Optimal Policies

Under the world model, a policy can be interpreted as a combination of parameters. Therefore, to find an optimal policy, two different types of information need to be provided. First, the criteria or the responses to be optimized need to be specified. Second, we need to determine the set of parameters we are able to and would like to control. Because including all parameters in this set will increase the complexity of the searching

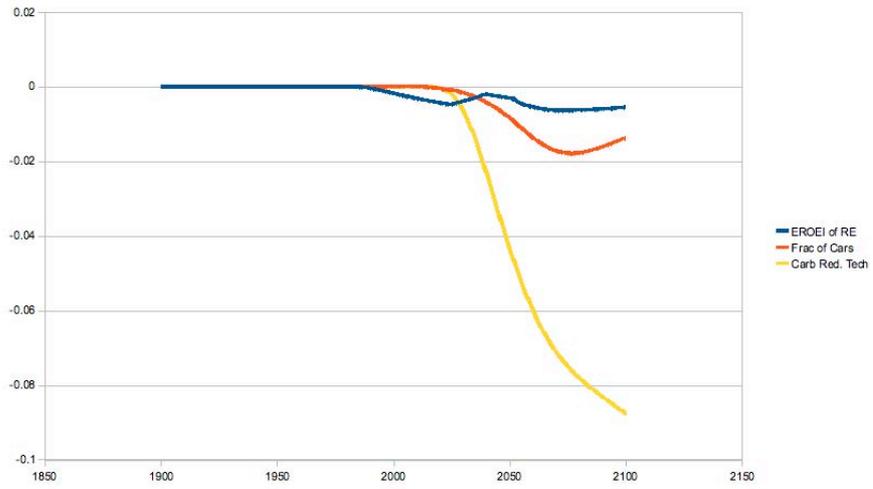


Figure 10: Comparison of Sensitivity of Carbon-dioxide to 3 parameters

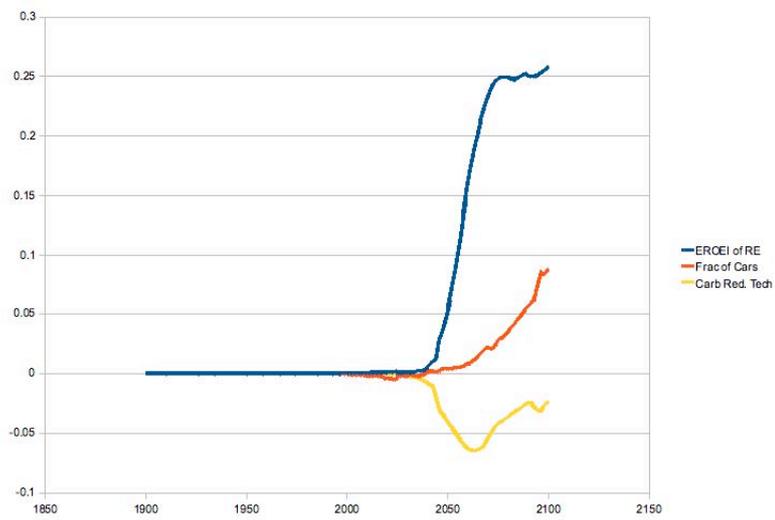
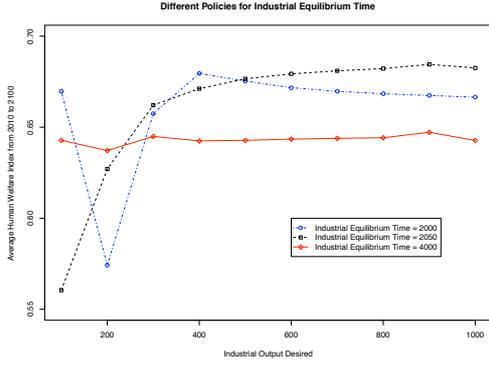
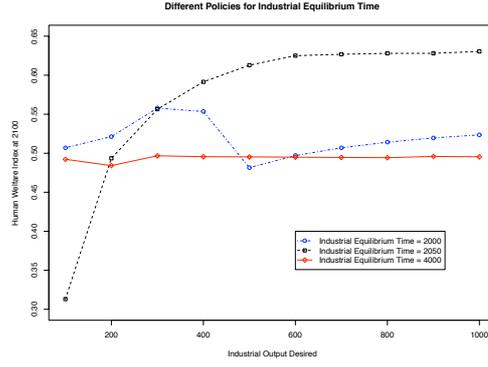


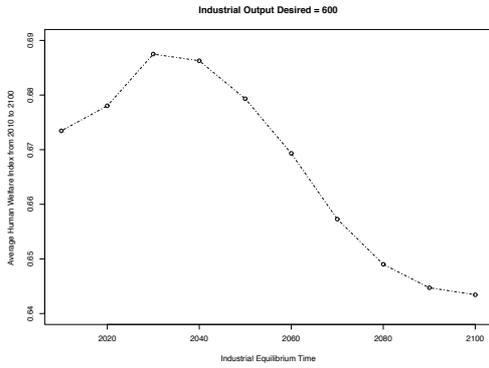
Figure 11: Comparison of Sensitivity of Carbon-dioxide to 3 parameters



(a) The effects of three industrial equilibrium times on Human Welfare Index averaged from 2010 to 2100.



(b) The effects of three industrial equilibrium times on Human Welfare Index at the end of the 21<sup>st</sup> century.



(c) The optimal industrial equilibrium time given the industrial output desired 600.

Figure 12: The optimal policy of Industrial Equilibrium Time and Industrial Output Desired for Human Welfare Index

problem, the above sensitivity analysis can be used first to find a smaller subset of parameters that has the most influence on the responses. After these sensitive parameters are selected, a steepest ascent gradient method [8] can be used to search for the optimal combination. Since the below results point out the possibility of local maxima, we may also need to restart the gradient method at many initial points so that we can obtain the global optimization.

We illustrate this policy searching problem through the problem of finding an optimal combination of Industrial Equilibrium Time and Industrial Output Desired per capita for Human Welfare Index. The example also describes the need of the world consensus on controlling the global industrial development; otherwise, the limited natural resources can not sustain the development of our society forever.

Since the free version of software Vensim does not allow us to setting up parameters and run simulations automatically, we will discuss here our manual method to search for such a optimal combination. One important issue is the outputs of the model are time dependent; therefore, we need to define how Human Welfare Index should be optimized. We consider two criteria: the average of Human Welfare Index from 2010 to 2100 and the Human Welfare Index at the end of the twentyfirst century, the year 2100.

The graphs 12(a) and 12(b) illustrate the effects of industrial output desired on Human Welfare Index under three different industrial equilibrium times. These graphs indicate that we need to set up the industrial equilibrium time somewhere between 2000 and 2100, and the industrial output desired at 600 so that Human Welfare Index will not only be improved but also last longer under the limitations of natural resources to sustain such development. The graph 12(c) point out the optimal time somewhere between 2030 to 2040 for industrial equilibrium given the industrial output desired 600.

## 7 Conclusions

There is much work to be done to probe this model for useful information. As we have stated, the model has some limitations, particularly in the simplicity of its assumptions with regard to renewable energy. Furthermore, there are issues surrounding water and land usage, sea-level, and changing location of farmland due to global climate change and their effects on food production. The impact of a possible shortage of phosphates, a necessary ingredient in fertilizers, should also be investigated. Given our limitations of time and software, there is much analysis that is yet to be done for this model. Our work suggests many avenues for further research. A fully functional version of Vensim would facilitate faster sensitivity analysis. The full version of Vensim provides many sensitivity analysis tools and allows for the running of Monte Carlo simulations which would significantly aid in the sensitivity analysis and the development of a reduced model.

The goal of future research should be to leverage the full abilities of Vensim and sensitivity analysis tools to add interactions to the model and then undertake model reduction of the World3-03 model.

Another issue unrelated to the implementation of the World3-03 model in Vensim is the challenge of finding accurate data on various parameters of the model. In particular, we could not find verified historical EROEI data for many resources. Research must be done to find accurate values of the EROEI of various resources from which we derive energy, particularly those that are inputs in renewable energy technologies.

We must remember the the World3-03 Model represents a broad description of various relationships that exist between human activity and the Earth's natural systems. It is not a model to predict the future. However, the model can be used to investigate the consequences of the complicated network of relations between finite resources, human activities, and the Earth's ecosystem.

## References

- [1] D. G. Cacuci, Mihaela Ionescu-Bujor, and Ionel Michael Navon. *Sensitivity and uncertainty analysis*. Chapman & Hall, Boca Raton, 2005.
- [2] Dolores García. A New World Model Including Energy and Climate Change Data. April 2009. <http://europe.theoildrum.com/node/5145>, accessed on July 26, 2009.
- [3] Energy Information Administration Official Energy Statistics from the U. S. Government 2009. <http://www.eia.doe.gov/emeu/international/contents.html>, accessed on July 24, 2009.
- [4] Thomas D. Kelly and Grecia R. Matos. Historical Statistics for Mineral and Material Commodities in the United States. 2009. <http://minerals.usgs.gov/ds/2005/140>, accessed on July 26, 2009.
- [5] Meadows, Donella H. and Richardson, John. and Bruckmann, Gerhart. *Groping in the Dark: the First Decade of Global Modelling*. Wiley, New York, 1982.
- [6] Donella H. Meadows, Jorgen Randers, and Dennis L. Meadows. *Limits to Growth: The 30-Year Update*. Chelsea Green, Vermont, 2004.
- [7] Robert Rapier. Predicting the Past: The Hubbert Linearization. March 2007. <http://www.theoildrum.com/node/2357>, accessed on July 26, 2009.
- [8] R. E. Steuer. *Multiple Criteria Optimization: Theory, Computation and Application*. Wiley, New York, 1986.
- [9] Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley, New York, 2002.

# Design of RF MEMS Switches without Pull-in Instability

Tyler Skorczewski<sup>1</sup>, Gregory P. Richards<sup>2</sup>, W. Cyrus Proctor<sup>3</sup>, Chongyi Shen<sup>4</sup>, Min Wang<sup>5</sup>, Jingyan Zhang<sup>6</sup>, Peng Zhong<sup>7</sup>, Ralph Smith<sup>8</sup>

Problem Presenter:

Jordan Massad

Sandia National Laboratories

## Abstract

*Micro-electro-mechanical systems (MEMS) switches for radio-frequency (RF) signals have certain advantages over solid-state switches, such as lower insertion loss, higher isolation, and lower static power dissipation. Mechanical dynamics can be a determining factor for the reliability of RF MEMS. The RF MEMS ohmic switch discussed in this paper consists of a plate, suspended over an actuation pad by four double-cantilever springs. Closing the switch with a simple step actuation voltage typically causes the plate to rebound from its electrical contacts. The rebound interrupts the signal continuity and degrades the performance, reliability and durability of the switch. The switching dynamics are complicated by a nonlinear, electrostatic pull-in instability that causes high accelerations. Slow actuation and tailored voltage control signals can mitigate switch bouncing and effects of the pull-in instability; however, slow switching speed and overly-complex input signals can significantly penalize overall system-level performance. Examination of a balanced and optimized alternative switching solution is sought. A step toward one solution is to consider a pull-in-free switch design. This project aims to determine how simple RC-circuit drive signals and particular structural properties influence the mechanical dynamics of an RF MEMS switch designed without a pull-in instability. The projects approach is to develop a validated modeling capability and subsequently study switch behavior for variable drive signals and switch design parameters. In support of project development, specifiable design parameters and constraints will be provided. Moreover, transient data of RF MEMS switches from laser Doppler velocimetry will be provided for model validation tasks. Analysis showed that a RF MEMS switch could feasibly be designed with a single pulse waveform and no pull-in instability and achieve comparable results to previous waveform designs. The switch design could reliably close in a timely manner, with small contact velocity, usually with little to no rebound even when considering manufacturing variability.*

*Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.*

## 1 Introduction and Motivation

Micro-electro-mechanical systems (MEMS)[6] have been used in various radio frequency (RF) applications to improve operation and performance [10][4][5][7]. RF MEMS ohmic switches have shown lower insertion loss, higher isolation, greater linearity, and lower static power dissipation than solid state switches [10]. Mechanical dynamics can be a determining factor for the reliability of RF MEMS. The RF MEMS ohmic switch discussed in this paper consists of a plate, suspended over an actuation pad by four double-cantilever springs; see Figure 1.1. During operation, an input voltage is typically applied to close an RF MEMS switch, and the dynamics of the closure event have a significant impact on the performance and design lifetime of a switch.

---

<sup>1</sup>University of California Davis

<sup>2</sup>Kent State University

<sup>3</sup>North Carolina State University

<sup>4</sup>University of Iowa

<sup>5</sup>Northern Illinois University

<sup>6</sup>Penn State University

<sup>7</sup>University of Tennessee Knoxville

<sup>8</sup>North Carolina State University

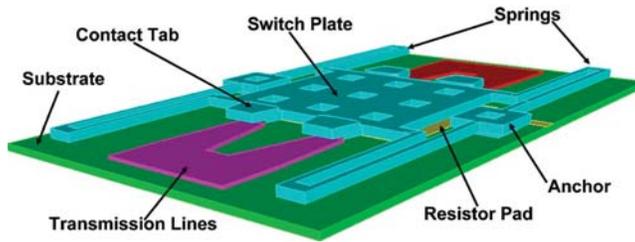


Figure 1.1: Three Dimensional RF-MEMS switch [2].

The kinetic energy of a switch scales with its velocity, and when a switch closes this energy must be dissipated before the switch attains a stable, closed state. This was observed experimentally using laser-Doppler vibrometry [9]; a switch was shown to bounce on its contacts a number of times before closing. The voltage signal used to close the switch was then replaced with a sequence of pulses that was shown to greatly reduce the velocity with which the switch impacted its electrical contacts and the subsequent bouncing. A high velocity before impact can also lead to higher deformations and, hence, higher stresses in the device during the close event. These observations motivate the use of shaped voltage waveforms to limit the contact velocities of electrostatically actuated MEMS devices.

There are many ways to design the waveform of the input voltage. In this paper, we focus on one pulse to close the switch and hold it in contact. Moreover, the work described in this paper seeks to model and optimize the dynamics of an RF MEMS switch while accounting for the unit-to-unit variability inherent in the fabrication of the switches.

## 2 Model Equations

Using standard reduction techniques, a single-degree-of-freedom model of the switch was constructed [2]. Most of the relevant physics of the free system could be captured using a driven damped harmonic oscillator as shown in Figure 2.1.

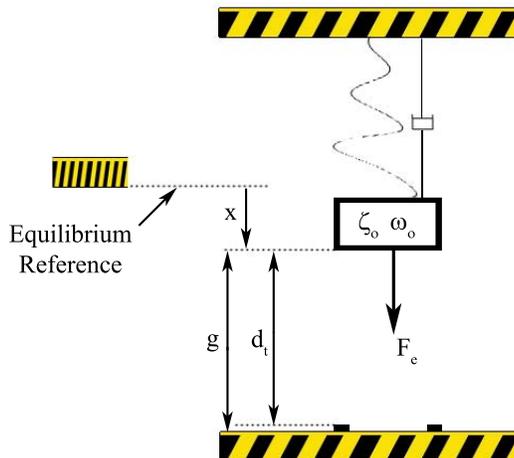


Figure 2.1: RF-MEMS switch model [2].

The equation for the model above is given as:

$$\ddot{x} + 2\zeta_o\omega_o\dot{x} + \omega_o^2x = \frac{F_e}{m_{eff}} \quad (2.1)$$

where  $m_{eff}$  is the effective mass of the plate-spring structure,  $\zeta_o$  is the dampening coefficient,  $\omega_o$  is the natural frequency of the spring system and  $F_e$  is the electrostatic force exerted on the plate defined by Eq.(2.2). From Figure 2.1,  $g$  is the distance between the plate and electrode at equilibrium,  $d_t$  is the travel distance of the switch until 'good contact' and  $x$  the position of the switch plate mass from equilibrium.

$$\frac{F_e}{m_{eff}} = \frac{\alpha V_h^2}{(g-x)^2} \quad (2.2)$$

with

$$\alpha = \frac{\epsilon A}{2m_{eff}}$$

where  $\epsilon$  is the permittivity of air,  $A$  is the cross-sectional area of the switch plate,  $V_h$  is the hold voltage. Substituting Eq.(2.2) into Eq.(2.1) yields

$$\ddot{x} + 2\zeta_o\omega_o\dot{x} + \omega_o^2x = \frac{\alpha V_h^2}{(g-x)^2} \quad (2.3)$$

Oscillation of the switch plate during contact in the closed position can approximately be captured by using the system shown in Figure 2.2. This secondary system was used only when the switch came in contact with the transmission line. The natural frequency and dampening coefficient,  $\omega_c$  and  $\zeta_c$ , of this model were fit using a least squares analysis of experimental data from closing of the switch. The impact model is given by

$$\ddot{x} + 2\zeta_c\omega_c\dot{x} + \omega_c^2x = \frac{\alpha V_h^2}{(g-x)^2} \quad (2.4)$$

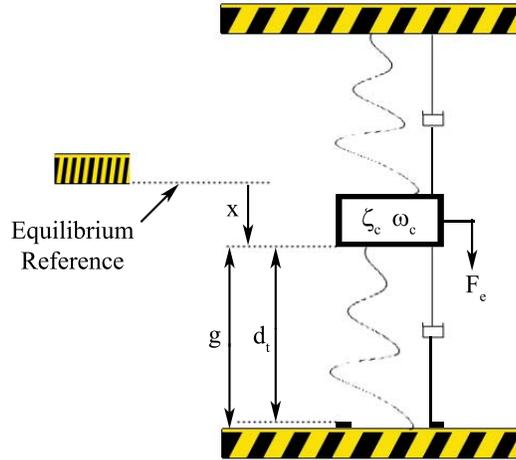
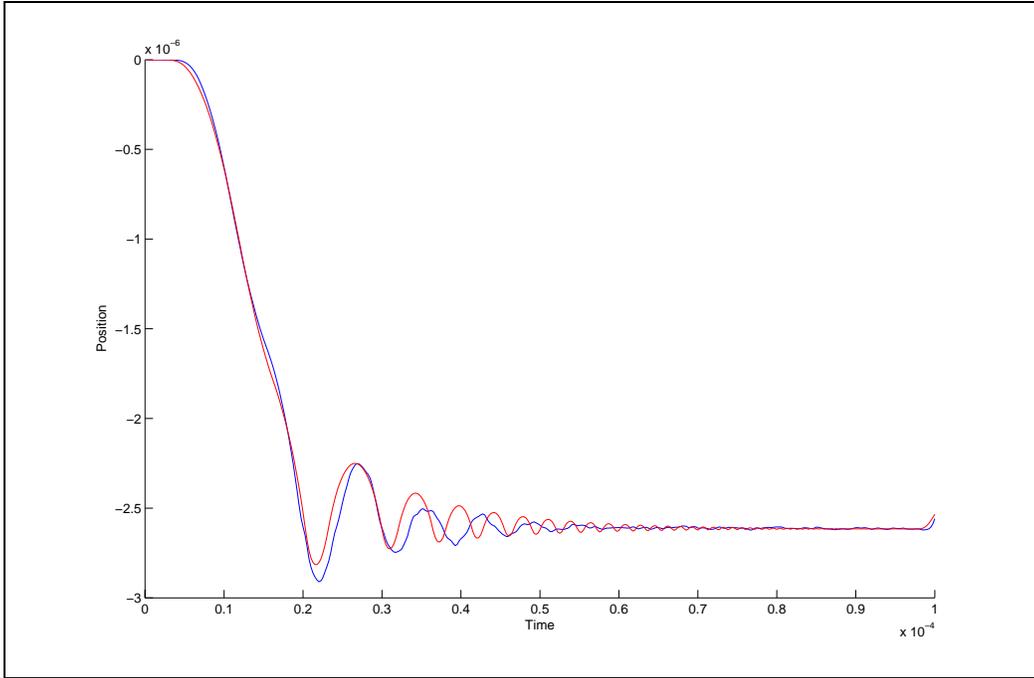


Figure 2.2: RF-MEMS switch model during impact [2].

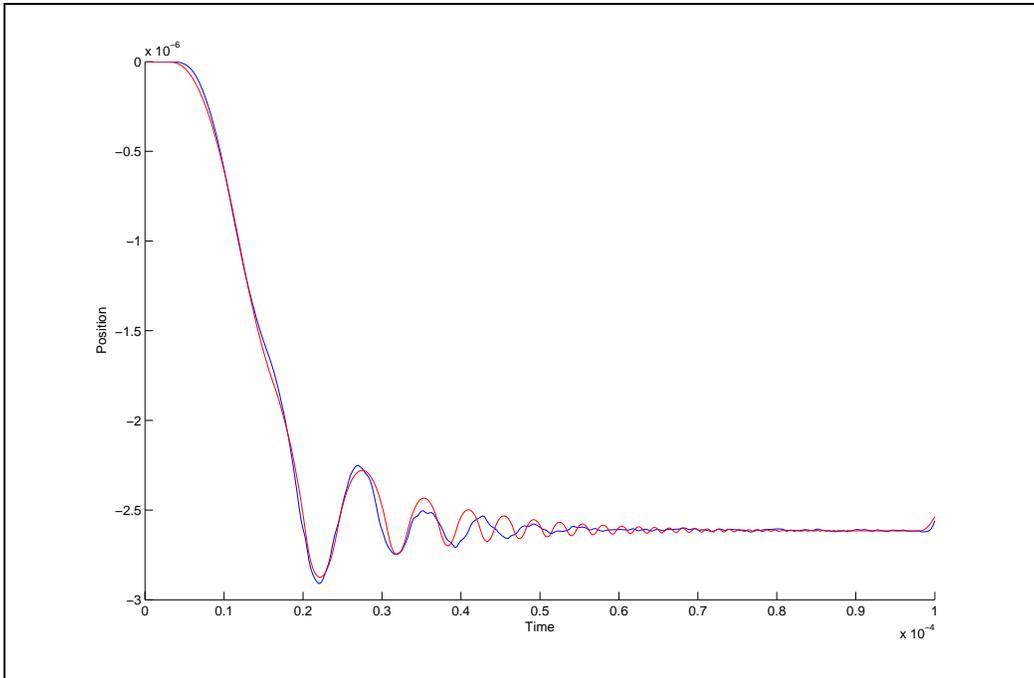
To fit  $\zeta_c$  and  $\omega_c$ , experimental data of position of the plate from equilibrium versus time was collected for multiple runs. Data for the fit was used while the plate was in contact with the electrodes. A time shift of  $0.75\mu s$  was used and data out to  $6.0\mu s$  was utilized for fitting purposes. Figure 2.3, illustrates before and after the fit. As the switch continues to bounce off the electrodes, higher-order effects may account for the discrepancy between the one dimensional model and the experimental data.

## 2.1 Pull-in Instability

A phenomenon known as pull-in instability can occur in electrostatic MEMS devices [3]. To capture the onset of instability we start by writing the total potential energy of the capacitive plate-spring system:



(a) Before fit.



(b) After fit.

Figure 2.3: Model calibration for closing switch data (a) before fit, (b) after fit.

$$E = -\frac{1}{2} \frac{\epsilon A}{(g-x)^2} V_h^2 + \frac{1}{2} K_{eff} x^2 \quad (2.5)$$

The first term represents the electrostatic potential energy of the variable plate capacitor and the second represents the potential energy stored in the harmonic oscillator. The force acting on the system then is given by

$$F = -\frac{\partial E}{\partial x} = \frac{1}{2} \frac{\epsilon A}{(g-x)^2} V_h^2 - K_{eff} x \quad (2.6)$$

The point of equilibrium occurs when the electrostatic force balances the spring restoring force and  $F = 0$ . The stiffness of the spring system can be obtained from the derivative of the force with respect to the distance from equilibrium

$$\frac{\partial F}{\partial x} = \frac{\epsilon A}{(g-x)^3} V_h^2 - K_{eff} \quad (2.7)$$

Then, using 2.6, we have

$$\frac{\partial F}{\partial x} = \frac{2K_{eff} x}{(g-x)} - K_{eff} \quad (2.8)$$

Solving for the inflection point of the instability

$$x = \frac{g}{3} \quad (2.9)$$

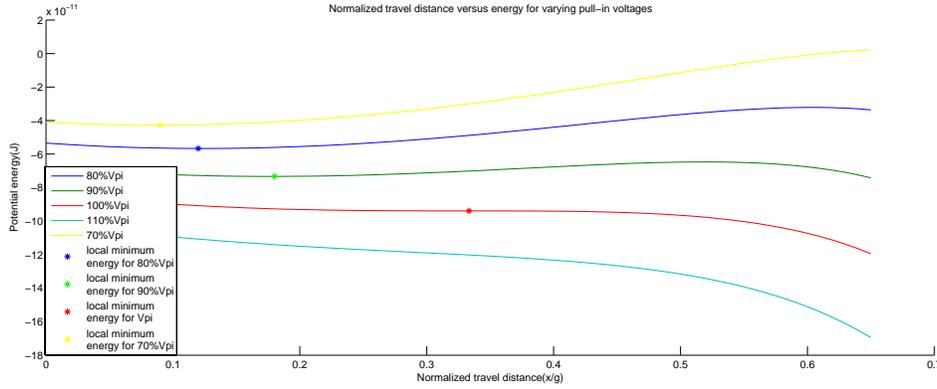


Figure 2.4: Pull-in instability: potential energy versus travel distance normalized by gap.

If the distance of travel,  $d_t$ , is less than  $1/3$  the gap distance no pull-in instability will occur. Beyond the inflection point, the plate will abruptly accelerate towards the resistor pad until the switch makes contact with the transmission lines. Figure 2.4 illustrates the onset of this instability. Note that the inflection point gradually increases until just beyond  $g/3$  it disappears all together. For this project, the travel distance will be chosen to ensure that the pull-in instability will never be reached.

### 3 Deterministic analysis

To minimize the contact velocities of the switch without a pull-in instability, we first consider a deterministic optimization for the design of the MEMS switch. This optimization will allow insight into the behavior of the system with a variety of travel distances and wave forms. Additionally, the optimizations along with the sensitivity analysis will give some intuition into which parameters are most important for running the larger uncertainty quantification problem. We describe our solution method and optimization procedure followed by a discussion some of the results of the deterministic optimization.

### 3.1 Optimization

With different values for  $\zeta$  and  $\omega$ , system contact is determined by

$$\zeta = \begin{cases} \zeta_o & x(t) < d_t \\ \zeta_c & x(t) \geq d_t \end{cases}$$

where  $d_t$  is the travel distance to contact and a similar relation for  $\omega$ .

To solve this ODE numerically, the second order ODE is first transformed into a system of first order ODE's via the standard method

$$\begin{aligned} \frac{dy_1}{dt} &= y_2 \\ \frac{dy_2}{dt} &= \frac{\alpha V_h(t)^2}{(g - y_1)^2} - 2\zeta\omega y_2 - \omega^2 y_1. \end{aligned}$$

Then two solvers are used, `ode45()` and `ode15s()`. The system can be compared to a damped oscillator, and `ode45()` can spend a long time finding exact values for small oscillatory behavior. Thus, when concerned with exactness for the entire behavior `ode45()` is employed. However, if we are only concerned about the behavior of the first few impacts with the contacts we use `ode15s()`. MATLAB's event locator is also used to determine the times of contact and the associated velocities.

We constrain the waveform of the applied voltage to one that can be easily achieved by a RC circuit, thus we provide the solvers with a voltage curve that can be determined by two parameters.

$$V_h(t) = \begin{cases} V_{hold} \left[ 1 - e^{-(t/t_b)} \right] & V_h(t) < 0.99 * V_{hold} \\ V_{hold} & \text{o.w.} \end{cases}$$

The next issue is to determine the metric to use for optimization. A variety of metrics are possible depending on which behavior is most desirable. Is a small contact velocity more important than a large time to closure of the switch? The metric that is used is a balance between the two. We minimize over energy in the system after first contact until the switch is considered closed. That is we minimize

$$F = \int_{t_{contact}}^{t_{close}} \dot{x}^2(t) dt.$$

$t_{close}$  is defined by the point in time in which the contact tabs on the switch plate come in 'good contact' with the transmission lines and stay within a defined tolerance. The point of 'good isolation' for the switch was defined by capacitance of 170e-18 Farads. Capacitance between two parallel plates serves as an upper bound when considering the electrostatic forces between the contact tabs and transmission lines. 'Good contact' was assumed to be 10 times less distance than the defined tolerance for 'good isolation'. Capacitance of an infinite parallel plate system is given as

$$C = \frac{\epsilon A_{ct}}{d} \quad (3.1)$$

where  $\epsilon = 8.8542\text{e-}12\text{F/m}$  is the permittivity of air,  $A_{ct}$  is the area of the contact tab surface and  $d$  is the distance between contact tab and waveguide. Contact tab areas may be machined down to  $1\mu\text{m}^2$  reliably which gives the absolute minimum distance before 'good isolation' occurs. This distance,  $d_{gi}$  is given as  $0.05208\mu\text{m}$ . The distance for 'good contact',  $d_{gc}$  is then  $0.005208\mu\text{m}$ . Note that this distance will affect the minimum travel distance allowed for the switch. If while the switch is oscillating freely the plate travels into this region that is smaller than 'good isolation' the switch may leak current.

To find the optimization of our objective function we use two of MATLAB's built in optimization tools, `fminsearch()` and `fminunc()`. `fminbnd()` is used for initial one dimensional investigations. Since `fminbnd()` is a combination of a golden section and Newton's method type search, it is restricted to one varying parameter. Thus, `fminsearch()` is used for optimization of the entire system. When using `fminsearch()` we do have to adjust our objective function slightly by adding in a penalty term for when the test parameters are outside of the physical bounds of the system.

We should note that `fminsearch()` is a Nelder-Mead Simplex direct search engine and is both highly dependent on the initial guess and has a tendency to get stuck in local minima when trying to optimize over several variables. To get around some of these limitations, we incorporate several one dimensional searches into the larger optimization, but we are not guaranteed that the global minimum is obtained.

We begin our deterministic optimization with several one dimensional searches using the `fminbnd()` routine. Our goal here is to find out if the possibility exists for a good switch design with no pull-in instability and a single pulse voltage curve. After setting the holding voltage to an arbitrary number of 95 volts, we optimize over the travel distance,  $d_t$ . The resulting position and velocity curves can be found in Figure 3.1.

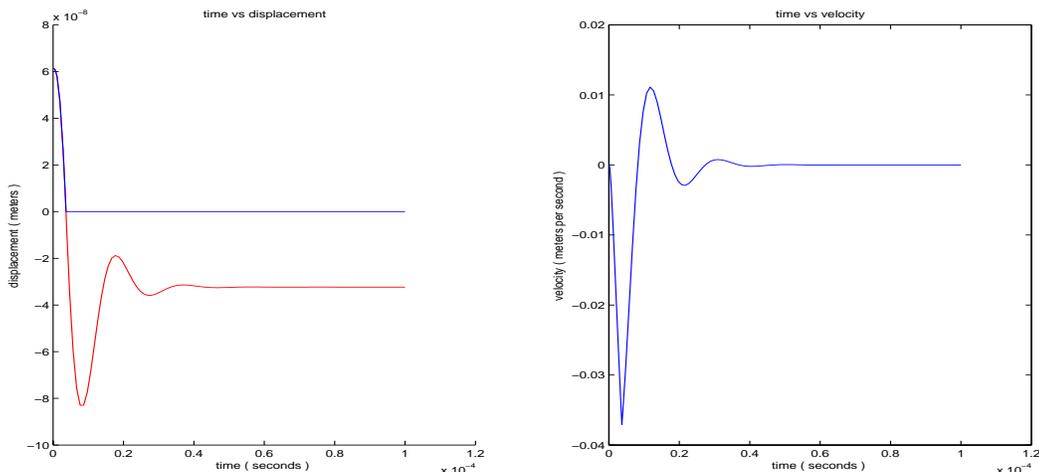


Figure 3.1: Example of an optimized position and velocity curve. Optimized over travel distance with a holding voltage at 95 volts, the red part of the position curve is when the switch is considered closed. As we can see, this curve has no residual bounces. The optimal travel distance was found to be  $6.13 \times 10^{-8}$  m.

From Figure 3.1, we see that a switch design that is activated using a single pulse waveform is possible. The switch design has several advantages. The time to closure of the switch is relatively short (roughly 4 microseconds), The system does not short out (that is the position curve does not go through the point  $x(t) = g$ ) and the switch has a low contact velocity of approximately 3.7 cm/s. However, this design does have some drawbacks. The travel distance is very close to a good separation distance where no signal can get through the switch. So, when the switch would be opened and oscillating back into equilibrium, the switch would spend a significant amount of time letting some signal through and thus could not be truly considered in the off position.

Once a travel distance is picked, one can then optimize over the waveform of the applied voltage. To keep the search one dimensional, we assume the build up time to be fixed at  $3.0 \times 10^{-7}$  s. Fixing the travel distance to be slightly less than one third of the gap distance ( $1.2 \times 10^{-6}$  m vs.  $3.8 \times 10^{-6}$  m), the optimization search produces a resulting position and velocity curves found in Figure 3.2 with an optimum voltage of roughly 148 V and an impact velocity of 27 cm/s.

Now we consider the results of several multi-dimensional searches over travel distance, gap distance, holding voltage, build up time, and spring stiffness ( $d_t, g, V_h, t_b$  and  $\omega_o$ ). Analyzing the results of several runs, we can

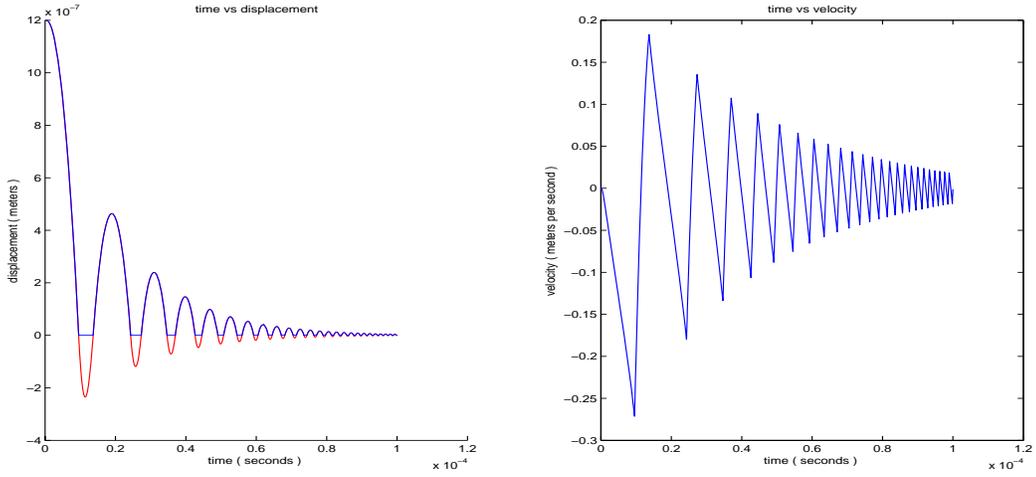


Figure 3.2: Example of an optimized position and velocity curve # 2. Optimized holding voltage over energy with a travel distance of  $1.2 \times 10^{-6}$  m, the red part of the position curve is when the switch is considered closed.

come to the realization that the search engine will always return an optimal travel distance that was equal to the minimum travel distance allowed. Thus we will optimize only over  $g$ ,  $V_h$ ,  $t_b$  and  $\omega_o$ . Some results of these searches are shown in Table 3.1.

Table 3.1: Deterministic optimization results. Optimization results setting a fixed travel distance,  $d_t$ , and using the total energy as the metric. Here  $g$  is the gap distance,  $V_h$  is the maximum voltage,  $t_b$  build up time,  $\omega_o$  is the natural frequency of the spring,  $cv$  is the contact velocity,  $t_{close}$  is the time to close the switch, and  $en$  is the energy from first contact to switch closure.

$d_t(m)$	$g(m)$	$V_h(V)$	$t_b(s)$	$\omega_o(Hz)$	$cv(m/s)$	$t_{close}(s)$	$en(m^2/s)$
1.2e-6	3.6e-6	149.1	3.0e-7	1.5e5	0.28	8.3e-5	3.1e-7
1.1e-6	3.3e-6	145.5	3.1e-7	1.6e5	0.30	5.9e-5	2.4e-7
9.7e-7	2.9e-6	131.1	3.4e-7	1.6e5	0.29	4.8e-5	1.9e-7
8.6e-7	2.8e-6	126.4	3.9e-7	1.6e5	0.27	4.5e-5	1.5e-7
7.4e-7	2.2e-6	85.9	4.1e-7	1.6e5	0.22	4.9e-5	1.1e-7
6.3e-7	1.9e-6	67.4	5.5e-7	1.6e5	0.18	4.6e-5	7.7e-8
5.1e-7	1.5e-6	51.3	7.4e-7	1.6e5	0.16	4.1e-5	5.1e-8
4.0e-7	1.5e-6	47.1	7.8e-7	1.6e5	0.12	4.4e-5	3.6e-8
2.8e-7	1.6e-6	45.1	7.8e-7	1.6e5	0.09	4.5e-5	2.1e-8
1.7e-7	1.7e-6	41.3	8.2e-7	1.6e5	0.06	3.6e-5	7.8e-9
6.3e-8	2.0e-6	30.4	1.1e-6	1.6e5	0.02	3.2e-5	1.1e-9

These results show several interesting facts about the behavior of the optimized system as a function of the travel distance. First, both the optimal hold voltage and the build up time to the hold voltage seem to have a linear relation to travel distance. Additionally, ( though with round off it is less clear in the table ) the stiffness of the spring appears to not affect the optimal design very much. The gap distance seems to have a linear relationship up to the travel distance of  $4.0 \times 10^{-7}$  m, but this is an artifact of the minimum gap distance.

Now that we have seen the results of our deterministic approach for the optimization of the switch, we will consider the sensitivity of our system with some of the input parameters. For this next goal, one set of base parameters is needed. We choose the travel distance of  $5.0 \times 10^{-7}$  m. For the optimal gap, we chose a secondary minimum with values found in Table 3.2 because the first optimization was very close to having a pull-in

instability.

### 3.2 Opening switch constraints

If we not only consider the closing dynamics of the MEMS switch, but also when the switch opens, we find new constraints that may need to be placed on the system. Here we are assuming that the closed switch is at rest and is opened by turning off the voltage and letting the system come to rest at its initial open condition. The equation governing these dynamics is

$$\begin{aligned} \ddot{x} + 2\zeta\omega_o\dot{x} + \omega_o^2x &= 0 \\ x(0) &= d_t, \quad \dot{x}(0) = 0 \end{aligned}$$

These equations are well known as the equations for unforced damped harmonic motion. The solution of these equations is

$$x(t) = e^{-\zeta\omega_o t} \left[ d_t \cos(\gamma t) + \frac{\zeta}{\sqrt{1-\zeta^2}} \sin(\gamma t) \right]$$

where  $\gamma = \omega_o\sqrt{1-\zeta^2}$  is the damped frequency of the system. The behavior of a typical representative switch is given in Figure 3.3. Notice the very strong oscillations after the switch has been reopened. Because of the typical values of natural frequency and dampening parameters of system these oscillations will not die out until well beyond  $100\mu\text{s}$ . Given that the switch will likely close again on the order of this time, there is need for more study on the effects of extreme displacement and/or velocity at the start of closing the switch again.

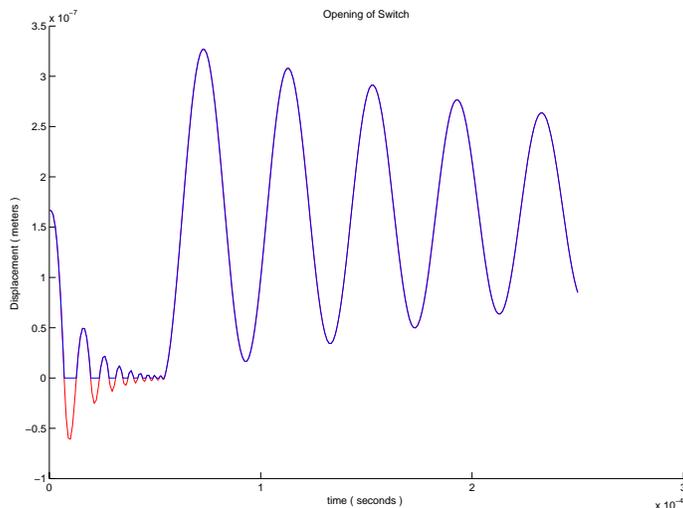


Figure 3.3: Shown here is a representative RF MEMS switch opening after having made contact. Because of the typical  $\omega$  and  $\zeta$  parameters the switch will oscillate strongly for a long period of time after opening.

Hypothetically, if we assume  $\zeta \ll 1$ , then we can say  $\frac{\zeta}{\sqrt{1-\zeta^2}}$  is negligible. In order to have the system relax to within 2% of the travel distance,  $d_t$ , in a specified time,  $\hat{t}$ , we then need

$$\begin{aligned}
x(\hat{t}) < 0.02d &\Rightarrow e^{-\zeta\omega_o\hat{t}} < 0.02 \\
&\Rightarrow -\zeta\omega_o\hat{t} < \log(0.02) \\
&\Rightarrow \omega_o > \frac{-\log(0.02)}{\zeta\hat{t}}
\end{aligned}$$

For  $\zeta = 0.02$  as in [2] and a relaxation time of  $\hat{t} = 10^{-4}$  seconds, this results in  $\omega \approx 850$  kHz. Since  $\omega_o = \sqrt{\frac{k_{eff}}{m_{eff}}}$ , if we assume the mass of the system is constant then we would need to require the spring constant to be approximately 50 times stronger than found in [2] [9] to have the switch open in a timely manner without any additional electrostatic force applied.

This itself poses new challenges, since a spring this strong is especially hard to pull on and would require a large voltage. A way to limit the need for large voltages is to shorten the travel distance,  $d_t$  and the gap between the resistor plate and the contact plate,  $g$ .

### 3.3 Sensitivity analysis

Both numerical and analytical sensitivity studies were carried out and compared. Below, each section is outlined and discussed. Comparisons of the sensitivities of the output responses to the chosen input parameters are analyzed from both studies.

#### 3.3.1 Numerical simulations

It is important to check how sensitive the computed solutions of the model are to the choice of input parameter values. Parameters used in the model of MEMS switch dynamics will vary during manufacturing, and any voltage waveform designed to close the switch should be effective over a range of values defined by the tolerances of fabrication. Before we undergo an uncertainty analysis of the model, we first must identify which parameters of the model are the most sensitive. Sensitivity here is checked through direct simulations where all parameters except the parameter of interest are held constant at an optimal value. Several computations are performed where the parameter of interest is varied across an interval determined from experience with manufacturing and previous simulations. These values are shown in Table 3.2. Extracted from the simulations are the time it takes to close the switch and the largest impact velocity the plate makes with the waveguides. One can infer from the relationships between these two variables and the parameter of interest how sensitive it is. Parameters with greater sensitivity should be considered in the uncertainty quantification study.

Table 3.2: Parameters investigated in sensitivity analysis

Parameter	Optimal Value	Min Value	Max Value
$d_t$ [m]	5.00e-07	5.00e-07	1.13e-06 ( $g_{opt}/3$ )
$g$ [m]	3.41e-06	1.50e-06	4.00e-04
$V_h$ [V]	1.27e+02	5.00e+01	1.5e+02
$t_b$ [s]	3.06e-07	1.00e-07	1.00e-04
$\omega_o$ [Hz]	1.49e+05	1.12e+05 (75% $\omega_o$ )	1.86e+05 (125% $\omega_o$ )
$x(0)$ [m]	0.00e+00	-1.00e-07 (-20% $d_{opt}$ )	1.00e-07 (+20% $d_{opt}$ )
$\dot{x}(0)$ [m/s]	0.00e+00	-1.55e-02 (-10% $vc_{opt}$ )	1.55e-02 (+10% $vc_{opt}$ )

The results of these simulations show that the relationships between impact velocity and several parameters are largely linear. See Figure 3.4. We can therefore quantify sensitivity as the magnitude of the slope of the line fitting a plot of scaled impact velocity and scaled parameter values.

$$S = \frac{\frac{\Delta R}{R_{opt}}}{\frac{\Delta \sigma}{\sigma_{opt}}} \quad (3.2)$$

Equation 3.2 outlines a generalized definition of sensitivity,  $S$ , of the normalized change in output response  $\Delta R = R - R_{opt}$  due to the change in normalized input  $\Delta\sigma = \sigma - \sigma_{opt}$ . These results are shown in table 3.3. From these results we see that impact velocity is most sensitive to changes in the gap thickness, peak voltage, initial displacement, and travel distance. It is moderately sensitive to natural frequency and not sensitive to changes in initial velocity or voltage ramp up time.

Table 3.3: Quantification of impact velocity sensitivity of model parameters

Parameter	Sensitivity
$d_t$	0.44
$g$	1.5
$V_h$	1.2
$t_b$	$\approx 0$
$\omega_o$	0.24
$x(0)$	0.51
$\dot{x}(0)$	0.075

Two parameters do not show a linear relationship with impact velocity. These are the initial velocity (quadratic) and the voltage ramp up time constant (exponential). See Figure 3.5. The initial velocity shows little sensitivity to impact velocity, and it takes more than an order of magnitude change in the ramp up constant to exhibit significant change in the impact velocity. It is interesting to note that impact velocity is minimized at an initial velocity greater than zero.

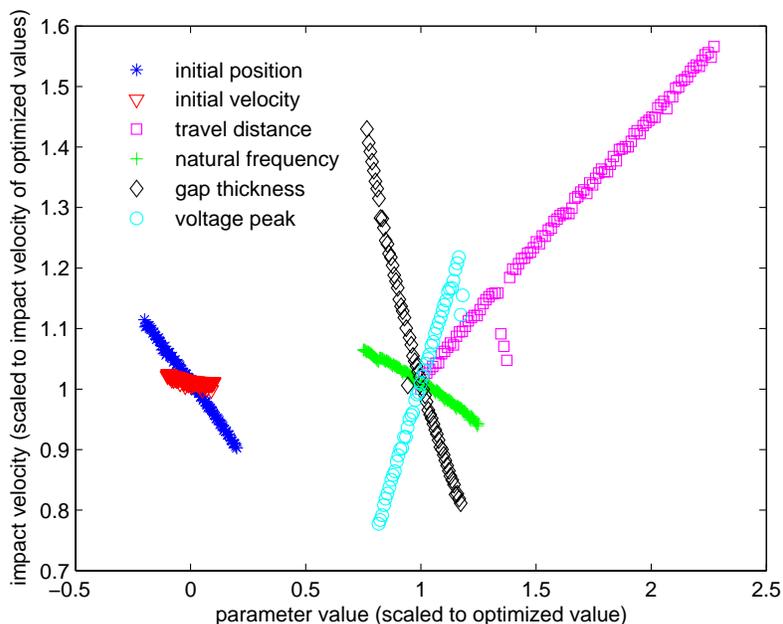


Figure 3.4: Sensitivity of model parameters on impact velocity. Shown are plots of impact velocity scaled to the impact velocity from a simulation with optimal parameter values versus parameter values scaled to optimized values. For initial velocity and initial displacement the parameter values are scaled to the impact velocity with optimized values and optimized travel distance, respectively. Larger slopes indicate more sensitivity.

The same analysis as above is also applied to the time it takes to close the switch. The results of the scaled closing time versus scaled parameter value plots again look linear, though this relationship is less well defined;

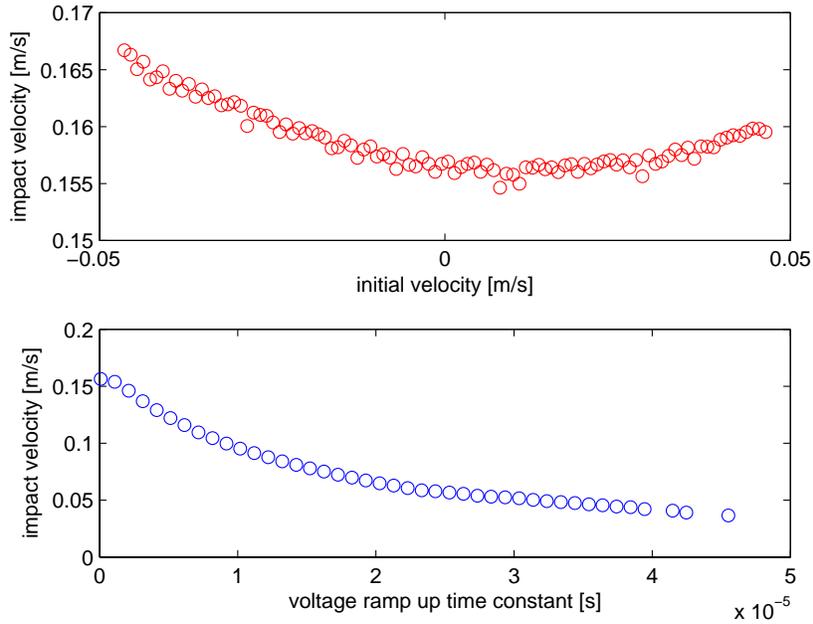


Figure 3.5: Shown here are plots of impact velocity versus initial velocity (top) and the voltage ramp up time constant (bottom). These two parameters do not display a linear relationship with impact velocity in the simulations of the RF MEMS switch. However, impact velocity is not sensitive to changes in either parameter.

see Figure 3.6. Again sensitivity is quantified as the magnitude of the slope of the linear fit to the data. The results shown in Table 3.4 show that switch closing time is most sensitive to peak voltage and gap thickness. These are followed by travel distance, natural frequency, and initial position. Closing time is least sensitive to initial velocity and the voltage ramp up time constant, as before.

Table 3.4: Quantification of switch closing time sensitivity of model parameters

Parameter	Sensitivity
$d_t$	0.64
$g$	2.8
$V_h$	2.9
$t_b$	$\approx 0$
$\omega_o$	0.49
$x(0)$	0.27
$\dot{x}(0)$	0.16

### 3.3.2 Analytic sensitivity analysis

Continuous sensitivity equation method (CSEM) is a common method to investigate the sensitivities of a system. Here we present the idea about analyzing the sensitivities of the one dimensional model with respect to the initial velocity  $\dot{x}(0)$ , initial displacement  $x(0)$ , natural frequency  $\omega$  and the gap distance  $g$ . Consider the model

$$\begin{cases} \ddot{x} + 2\zeta\omega\dot{x} + \omega^2x = \frac{\alpha V_h^2}{(g-x)^2}, \\ x(0) = a, \dot{x}(0) = b \end{cases} \quad (3.3)$$

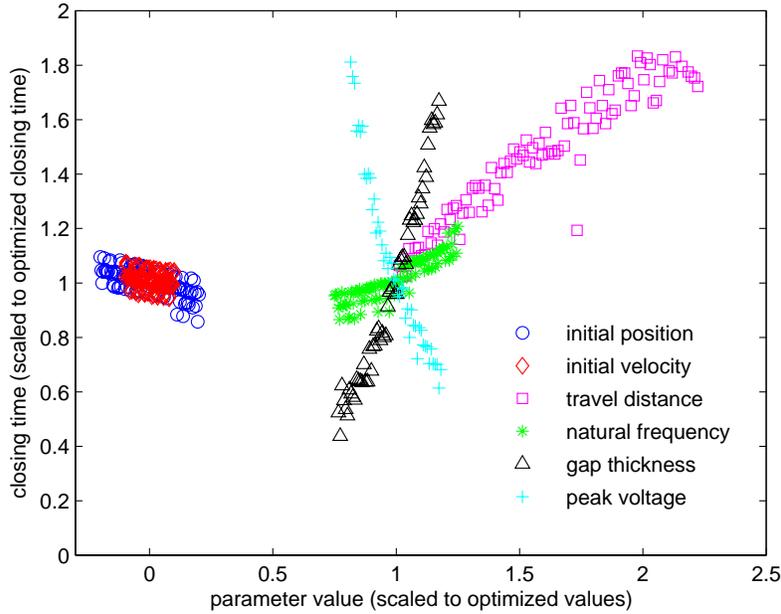


Figure 3.6: Sensitivity of model parameters on switch closing time. Shown are plots of closing time scaled to the closing time from a simulation with optimal parameter values versus parameter values scaled to optimized values. For initial velocity and initial displacement the parameter values are scaled to the impact velocity with optimized values and optimized travel distance, respectively. Larger slopes indicate more sensitivity.

where

$$\alpha = \frac{\varepsilon A}{2m_{eff}}, \quad \omega = \sqrt{\frac{K_{eff}}{m_{eff}}}$$

Let  $\omega_0$  and  $g_0$  be the optimized parameter values discussed previously. Sensitivity equations are used to investigate the rate of change of the output response of a system with respect to the input parameters. For a more comprehensive overview, see [8].

If we fix every parameter except  $b$ , then the solution  $x$  of (3.3) is a function that depends on  $t$  and  $b$ , i.e.,

$$x = x(t, b). \quad (3.4)$$

Assume that  $x$  is  $C^3$  on  $b$ . Let  $y_1 = \frac{\partial x}{\partial b}$  be the sensitivity of  $x$  with respect to  $b$ . Then

$$y_1' = \frac{\partial}{\partial t} y_1 = \frac{\partial^2 x}{\partial t \partial b} = \frac{\partial}{\partial b} \dot{x}. \quad (3.5)$$

It is clear that  $y_1$  represents the rate of change in displacement  $x$  to the initial velocity  $b$ . From (3.5),  $y_1'$  represents the rate of change in velocity  $\dot{x}$  to  $b$ . By taking the derivative of (3.3) to  $b$ , the following sensitivity equation can be derived

$$\begin{cases} y_1'' + 2\zeta\omega y_1' + (\omega^2 - \frac{2\alpha V_h^2}{(g-x)^3})y_1 = 0, \\ y_1(0) = 0, \quad y_1'(0) = 1. \end{cases} \quad (3.6)$$

Substituting (3.4) into (3.6) and solving, we obtain a solution  $y_1 = y_1(t, b)$  and hence  $y_1' = y_1'(t, b)$ . Let  $\dot{x}(t_1, b)$  be the first contact velocity of (3.3), and  $\dot{x}(t_2, 0)$  be the first contact velocity of (3.3) with  $b = 0$ , where  $t_1 = t_1(b)$  and  $t_2$  are the first contact time for each case respectively. Then, we have following estimations

$$\begin{aligned}
|\dot{x}(t_1, b) - \dot{x}(t_2, 0)| &= |\dot{x}(t_1, b) - \dot{x}(t_1, 0) + \dot{x}(t_1, 0) - \dot{x}(t_2, 0)| \\
&\leq |\dot{x}(t_1, b) - \dot{x}(t_1, 0)| + |\dot{x}(t_1, 0) - \dot{x}(t_2, 0)| \\
&\leq |y'_1(t_1, 0)||b| + |\dot{x}(t_1, 0) - \dot{x}(t_2, 0)|,
\end{aligned} \tag{3.7}$$

and

$$\begin{aligned}
\dot{x}(t_1, b) - \dot{x}(t_2, 0) &= \dot{x}(t_1, b) + \dot{x}(t_1, 0) - \dot{x}(t_1, 0) - \dot{x}(t_2, 0) \\
&\simeq y'_1(t_1, 0)b + \dot{x}(t_1, 0) - \dot{x}(t_2, 0),
\end{aligned} \tag{3.8}$$

$$= y'_1(t_1(b), 0)b + \dot{x}(t_1(b), 0) - \dot{x}(t_2, 0), \tag{3.9}$$

where  $y'_1(t, 0)$  and  $\dot{x}(t, 0)$  are known functions.

Let  $y_2 = \frac{\partial x}{\partial a}$  represent the sensitivity of the displacement,  $x$ , to initial displacement,  $a$ . The following sensitivity equation be derived from Eq.( 3.3)

$$\begin{cases} y''_2 + 2\zeta\omega y'_2 + (\omega^2 - \frac{2\alpha V_h^2}{(g-x)^3})y_2 = 0, \\ y_2(0) = 1, y'_2(0) = 0, \end{cases} \tag{3.10}$$

Where  $y'_2$  represents the rate of change in velocity to  $a$ . Using a similar method to calculating the sensitivity to  $b$ , the following estimations are obtained

$$\begin{aligned}
|\dot{x}(t_1, a) - \dot{x}(t_2, 0)| &= |\dot{x}(t_1, a) - \dot{x}(t_1, 0) + \dot{x}(t_1, 0) - \dot{x}(t_2, 0)| \\
&\leq |\dot{x}(t_1, a) - \dot{x}(t_1, 0)| + |\dot{x}(t_1, 0) - \dot{x}(t_2, 0)| \\
&\leq |y'_2(t_1, 0)||a| + |\dot{x}(t_1, 0) - \dot{x}(t_2, 0)|,
\end{aligned} \tag{3.11}$$

and

$$\begin{aligned}
\dot{x}(t_1, a) - \dot{x}(t_2, 0) &= \dot{x}(t_1, a) + \dot{x}(t_1, 0) - \dot{x}(t_1, 0) - \dot{x}(t_2, 0) \\
&\simeq y'_2(t_1, 0)a + \dot{x}(t_1, 0) - \dot{x}(t_2, 0).
\end{aligned} \tag{3.12}$$

Note that  $t_1$  in this case may not be the same as  $t_1$  with regard to the other sensitivity variables.

Let  $y_3 = \frac{\partial x}{\partial \omega}$  be the sensitivity of  $x$  to  $\omega$ , we can derive the following sensitivity equation from (3.3). The estimations are

$$\begin{cases} y''_3 + 2\zeta\omega y'_3 + (\omega^2 - 2\alpha V_h^2/(g-x)^3)y_3 + 2\zeta\dot{x} + 2\omega x = 0, \\ y_3(0) = y'_3(0) = 0. \end{cases} \tag{3.13}$$

$$\begin{aligned}
|\dot{x}(t_1, \omega) - \dot{x}(t_2, \omega_0)| &= |\dot{x}(t_1, \omega) - \dot{x}(t_1, \omega_0) + \dot{x}(t_1, \omega_0) - \dot{x}(t_2, \omega_0)| \\
&\leq |\dot{x}(t_1, \omega) - \dot{x}(t_1, \omega_0)| + |\dot{x}(t_1, \omega_0) - \dot{x}(t_2, \omega_0)| \\
&\leq |y'_3(t_1, \omega_0)||\omega - \omega_0| + |\dot{x}(t_1, \omega_0) - \dot{x}(t_2, \omega_0)|,
\end{aligned} \tag{3.14}$$

and

$$\begin{aligned}
\dot{x}(t_1, \omega) - \dot{x}(t_2, \omega_0) &= \dot{x}(t_1, \omega) + \dot{x}(t_1, \omega_0) - \dot{x}(t_1, \omega_0) - \dot{x}(t_2, \omega_0) \\
&\simeq y'_3(t_1, \omega_0)(\omega - \omega_0) + \dot{x}(t_1, \omega_0) - \dot{x}(t_2, \omega_0).
\end{aligned} \tag{3.15}$$

Let  $y_4 = \frac{\partial x}{\partial g}$  be the sensitivity of  $x$  to  $g$ . We can derive the following results

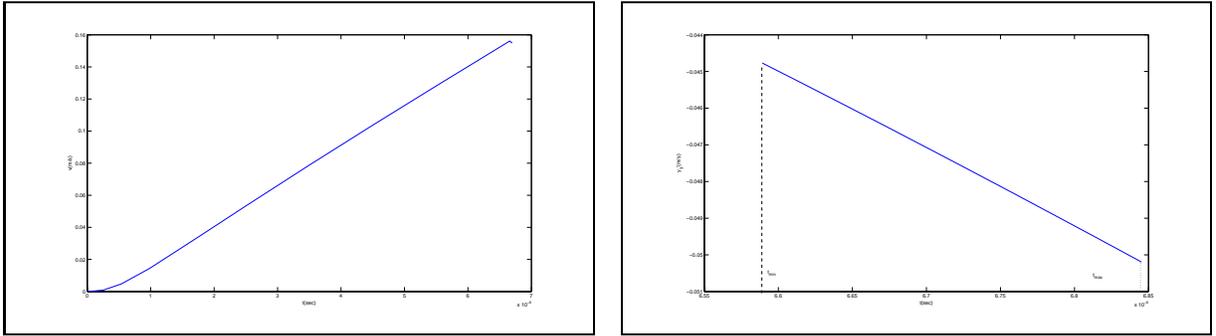
$$\begin{cases} y_4'' + 2\zeta\omega y_4' + (\omega^2 - 2\alpha V_h^2(g-x)^{-3})y_4 + 2\alpha V_h^2/(g-x)^3 = 0, \\ y_4(0) = 0, y_4' = 0. \end{cases} \quad (3.16)$$

$$\begin{aligned} |\dot{x}(t_1, g) - \dot{x}(t_2, g_0)| &= |\dot{x}(t_1, g) - \dot{x}(t_1, g_0) + \dot{x}(t_1, g_0) - \dot{x}(t_2, g_0)| \\ &\leq |\dot{x}(t_1, g) - \dot{x}(t_1, g_0)| + |\dot{x}(t_1, g_0) - \dot{x}(t_2, g_0)| \\ &\leq |y_4'(t_1, g_0)||g - g_0| + |\dot{x}(t_1, g_0) - \dot{x}(t_2, g_0)|, \end{aligned} \quad (3.17)$$

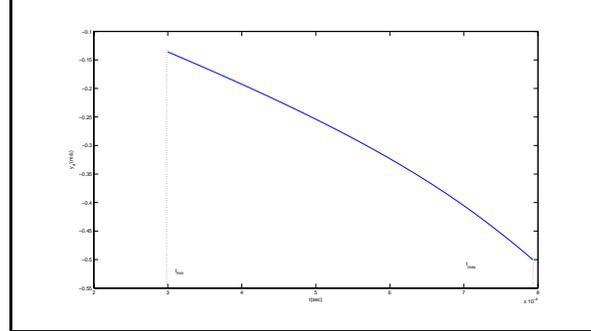
and

$$\begin{aligned} \dot{x}(t_1, g) - \dot{x}(t_2, g_0) &= \dot{x}(t_1, g) + \dot{x}(t_1, g_0) - \dot{x}(t_1, g_0) - \dot{x}(t_2, g_0) \\ &\simeq y_4'(t_1, g_0)(g - g_0) + \dot{x}(t_1, g_0) - \dot{x}(t_2, g_0). \end{aligned} \quad (3.18)$$

We solve (3.3) with optimized parameters, then solve (3.13) and (3.16). Numerical solutions  $x'(t, 0)$ ,  $y_3'(t, \omega_0)$  and  $y_4'(t, g_0)$  are obtained as Figs 3(a), 3(b) and 3(c). Those solutions enable us to estimate the rate of change of first contact velocity with respect to  $\omega$  and  $g$ .



(a) Numerical solution of  $x'(t, 0)$ . The velocity of (3.3) with optimized parameters. (b) Numerical solution of  $y_3'(t, \omega_0)$ . The rate of change of contact velocity to  $\omega$  around  $t_2$  at  $\omega_0$ .

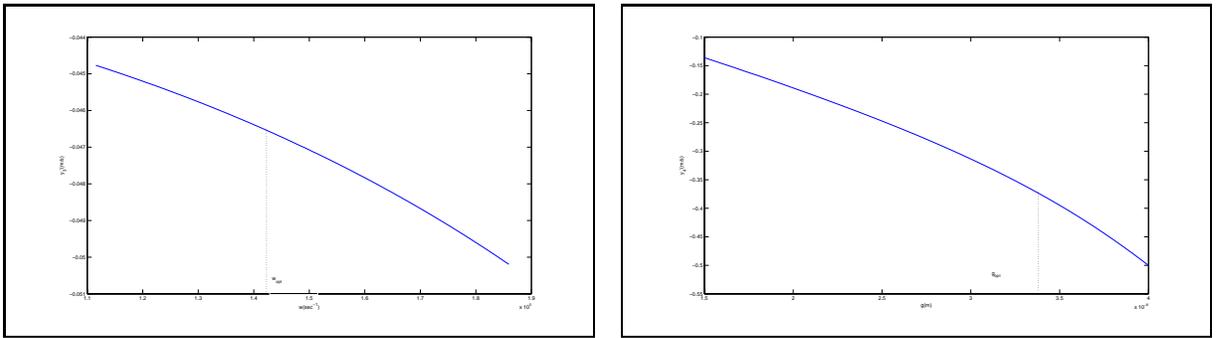


(c) Numerical solution of  $y_4'(t, g_0)$ . The rate of change of contact velocity to  $g_0$  around  $t_2$  at  $g_0$ .

Figure 3.7: Numerical estimations of values needed to solve for the bounds in equation 3.14 and 3.17. These allow for the solutions of normalized sensitivities of  $g$  and  $\omega$ .

Since the first contact time  $t_1$  depends on  $\omega$  and  $g$  respectively, we obtain a first contact time interval about  $t_1$  by numerical simulation. To compare the sensitivities of them, we need to normalize  $y_3'$  and  $y_4'$ . Let  $\hat{y}_3' = y_3'\omega_0$  and  $\hat{y}_4'g_0$  be the normalized sensitivities. From the Figures 3(a) 3(b), we know that (3.3) is more sensitive to  $g$  than  $\omega$ .

If we can obtain the analytic solution of (3.3), then the continuous sensitivity equation method can be applied to (3.3) to investigate its sensitivities to different input parameters. This will be an interesting topic because (3.7), (3.11), (3.14) and (3.17) show the upper bounds of the differences of contact velocities while (3.8), (3.12),



(a) Numerical solution of  $\hat{y}'_3$ . The rate of change of normalized sensitivity  $y'_3$  to  $\omega$ . (b) Numerical solution of  $\hat{y}'_4$ . The rate of change of normalized sensitivity  $y'_3$  to  $g$ .

Figure 3.8: Normalized sensitivities of contact velocities with respect to input parameters  $\omega$  and  $g$ .

(3.15) and (3.18) present a way to approximate the differences of contact velocities. These approximations will be very helpful to understand (3.3). More work can be carried out to find relationships between the sensitivities of the initial displacement and initial velocity for the analytic and numeric sections.

Due to lack of time, more analysis is needed to fully understand the behavior of the sensitivity equations when compared with the numerical results. One would expect that each should verify the other with approximations and error taken into account.

## 4 Uncertainty Quantification

The deterministic model defined previously minimizes over a form of energy which takes both contact velocities and closing times into consideration. The optimization procedure suggests the values of parameters, such as  $g$  and  $d_t$ , for the design of the switch. First, uncertainty quantification is a method to enhance the robustness of the optimization due to variations in fabrication and use. Second, UQ seeks an optimized distribution of  $g$ ,  $d_t$  and  $x_0$  for the manufacturer to set target goals in mass production.

### 4.1 Waveform optimization

From design optimization, we come up with a set of output parameters that are considered to be optimized. However, these parameters are for one single switch. While in mass production, one must also consider random unit-to-unit variability between the switches due to manufacturing techniques. Namely, it is that even if we target the gap at  $3.5\mu m$  for every single switch (for example), their actual gaps may be  $3.4\mu m$ ,  $3.6\mu m$ , and so forth. The same variability could apply to travel distance,  $K_{eff}$ , and  $m_{eff}$ .

Our existing data of past experiments and relevant analysis suggest that a waveform designed to optimize one switch cannot give good results when applied to an ensemble of switches, say, up to 1,000 variations. This is because of the existence of variability mentioned above and because these statistics have not been incorporated in the design. The uncertainty analysis aims to provide a waveform to an ensemble of switches so that the overall performance is optimized. This waveform can be potentially different than the one designed for one specific switch.

From the sensitivity analysis, we notice that  $\omega$  is not very sensitive to our optimization. In other words, a change of  $\omega$  does not impact the output significantly. As such, we would exclude  $\omega$  in the uncertainty analysis. Similarly, we will not consider the variability of  $K_{eff}$  and  $m_{eff}$ , which both are linked to  $\omega$ . Therefore what we will vary at this time is initial displacement, gap and travel distance.

The distribution of these random variables should be specified. Previous experimental measurements indicated that generalized Beta distribution is appropriate and conservative for both  $g$  and  $d_t$  but their distribution parameters are yet to be estimated. For initial travel, we would assume that it is uniformly distributed. A code has been developed to plot generalized Beta distribution with different  $\alpha$ (skewness),  $\beta$ (kurtosis), lower bound and upper bound, so that they can be compared with the plot we have from past experiments [1] to get closest estimates of  $\alpha$  and  $\beta$ . Our final estimates of  $\alpha$  and  $\beta$  for  $g$  are 2.0 and 4.8 respectively; for  $d_t$  they are 2.0 and 4.5 respectively.

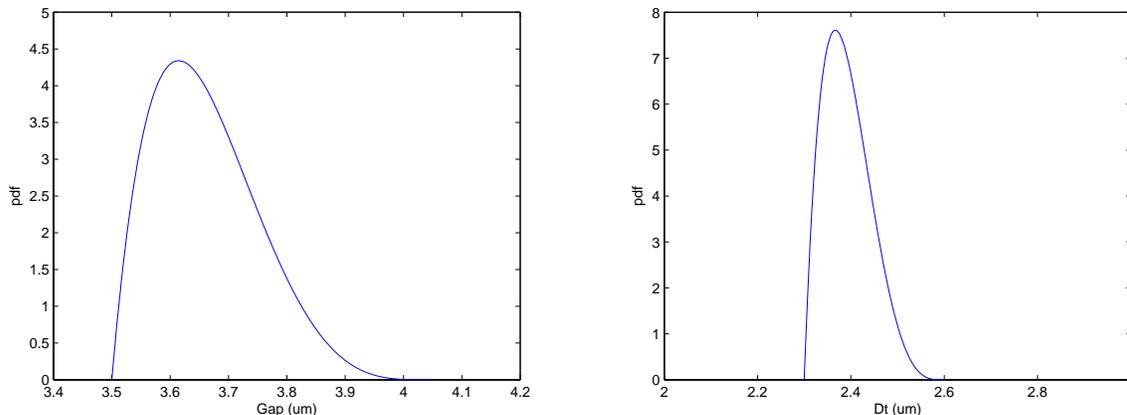


Figure 4.1: Distribution of gap and  $d_t$  after parameter estimation.

Still using the same computational routine, we took 25 samples from each of the parameter distributions, which form a total of  $25^3 = 15625$  simulations. The three variables are gap, travel distance, and initial displacement. We will first take a look at the contact velocity histogram without considering unit-to-unit variability (which will be called the 'deterministic model' hereafter). As for initial travel, we assume there are three scenarios (low, medium, and high travel distance values).

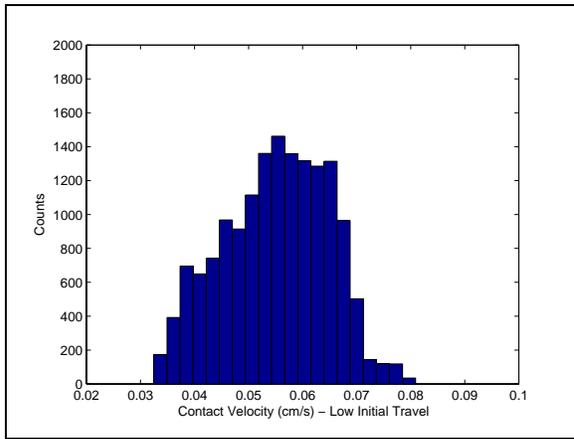
With this set of parameters we establish a base to compare with later. When initial travel is medium, up to  $6967/15625 = 45\%$  of the switches failed. This is due to the medium initial travel is close to the pull-in instability point we discussed previously. When initial travel is high, almost no switch can close with a speed less than  $25\text{cm/s}$ . We present this result only to illustrate the problem when we ignore the unit-to-unit variability among the devices, and for a comparison purpose how better we can make with including random variability.

Next, we will admit the existence of unit-to-unit variability, and design the waveform so that optimized overall performance is achieved. The optimization process we used is to minimize the 90th percentile of contact velocity  $v_{90}$ , which is defined by:

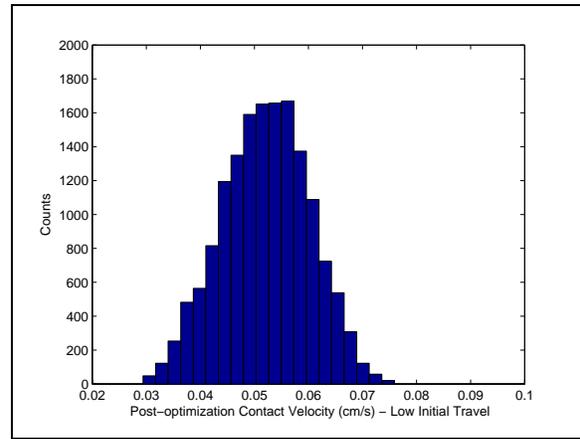
$$P(v > v_{90}) = 0.1.$$

Still initial displacements are assigned to the three low, medium and high scenarios; we will see how the frequency distribution of contact velocity changes with this optimization; see 4.2

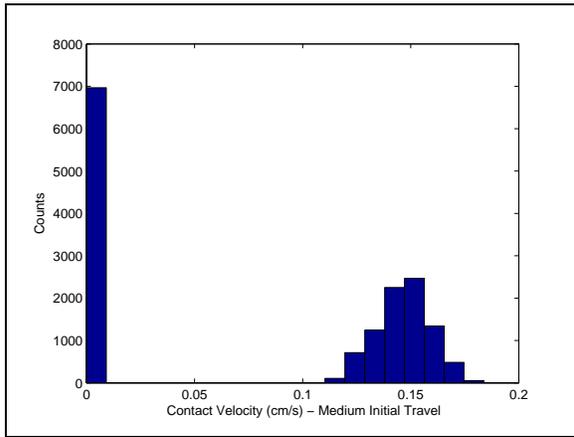
Result: improvement is obvious. When medium travel is assumed, the switch fail rate dropped rapidly from 45% to about 15%. While initial travel is high, even from visual comparison, one can easily perceive the effect of optimization. From a statistical perspective, it is also not hard to show that the improvement observed is quite significant. To be specific, let's calculate an asymptotic 95% confidence interval for the deterministic model's contact velocities and also for the probability models. Since the distribution of contact velocity is not known, we use  $\bar{v} \pm 1.96 \times \sqrt{Sv^2/N}$  to calculate the asymptotic 95% interval. As such, we get the interval for deterministic model is (0.2521, 0.3119), while for the probability model is (0.1752, 0.2402). These ranges have absolutely no overlap with each other. This is strong evidence that an optimization is necessary when we consider the parameter to be variable rather than 'theoretically fixed'.



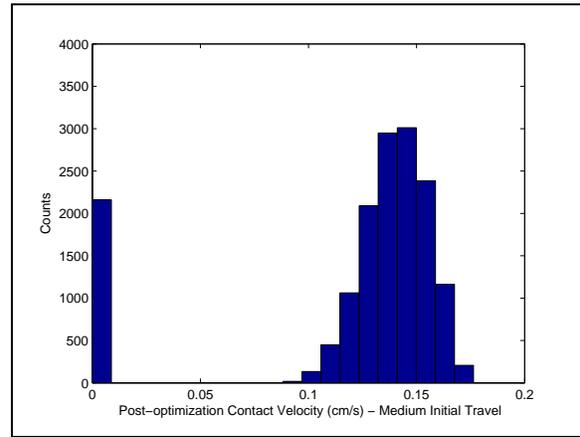
(a) Optimized low travel closing times.



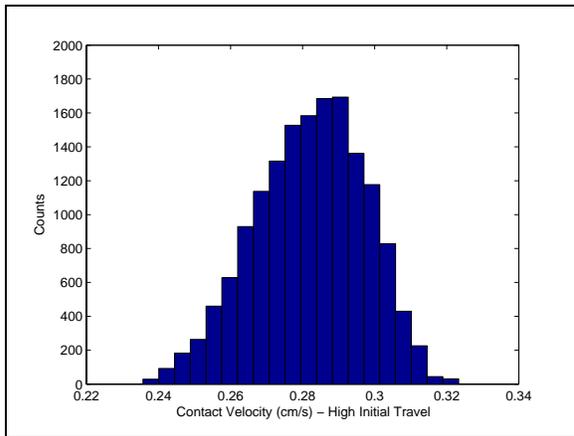
(b) Uncertainty optimized low travel energies.



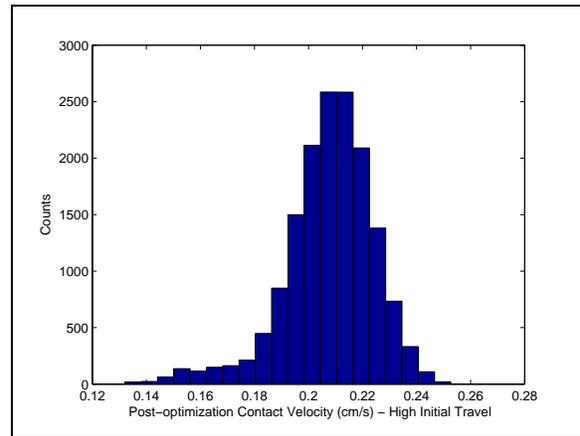
(c) Optimized medium travel contact velocities.



(d) Uncertainty optimized medium contact closing times.



(e) Optimized high travel energies.



(f) Uncertainty optimized high travel contact velocities.

Figure 4.2: Comparison of optimized values before and after uncertainty analysis using contact velocity as objective function.

Table 4.1: Comparison of Waveform Designs Between Deterministic Model and Uncertainty Model.

<b>Initial Travel:</b>	$V_h(\text{Deter})$	$t_b(\text{Deter})$	$V_h(\text{Uncer})$	$t_b(\text{Uncer})$	<b>Initial <math>E(g)</math></b>	<b>Initial <math>E(d_t)</math></b>
Low	41.303	8.1655e-7	38.403	8.4693e-7	1.31e-6	1.668e-7
Medium	51.309	7.4383e-7	49.385	7.9032e-7	1.5338e-6	5.112e-7
High	149.11	3.0068e-7	124.95	3.3256e-7	3.6037e-6	1.2e-6

## 4.2 Design optimization

The following work studies how the change of parameter distribution effects the optimized result. We start from shifting the support of the distributions of parameters and investigate the generated pdf histograms of optimized contact velocity, closing time and energy. (More work could be done in the future by changing other parameters, such as skewness and kurtosis of the Beta distributions.) Following each distribution of  $g$ ,  $d_t$ , and  $x_0$ , 10 random values are generated for each parameter, respectively, giving us totally  $1,000 = 10^3$  samples. Thus, we get a corresponding distribution set of contact velocities, closing times, and energies from those 1000 samples. The goal is to pick up a shift value with provides "best" optimized distribution set.

### 4.2.1 Performance Metrics

The objective function to be minimized is set up as

$$J(\epsilon) = V_p + c_1 p_f,$$

where  $\epsilon$  denotes the shift of distribution set,  $V_p$  denotes the 90 percentile of the optimized contact velocity distribution,  $p_f$  denotes the probability that the switch fails to close, and  $c_1$  is a weight to balance the energy and the number of switches which fails to close [2].

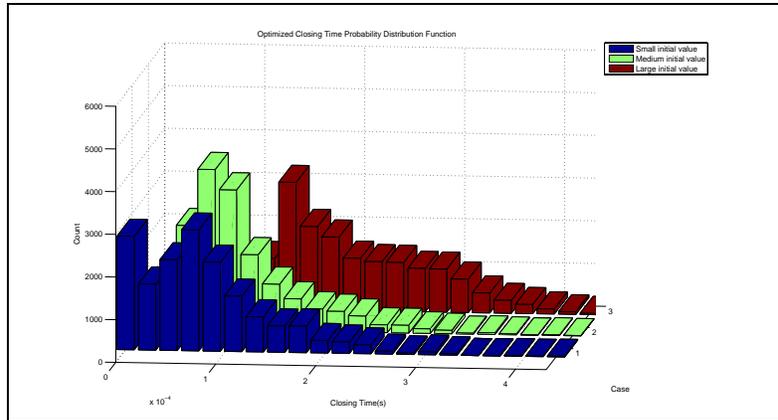
### 4.2.2 Computational Routine and Results

In the computational routine, the shift of distribution is characterized by the variation of expectation value. Since the deterministic model provides several sets of optimal parameters, the uncertainty analysis is done within three different parameter domains distinguished by the value of travel distance  $d_t$ . We elect three different  $d_t$  values, one relatively small, one medium, and one relatively large, to be the initial expectation values of travel distance distributions. We do separate iterations for each to obtain three sets of optimized results. Due to the time limit, all the numbers of iterations are fixed, thus the results shown in this paper are not the best that we could achieve.

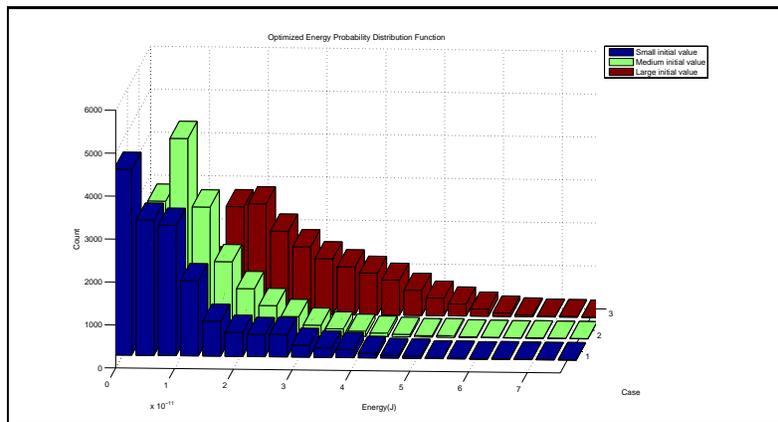
The histograms for optimized distribution of contact velocity, energy and closing time are given as in Figure 4.3, which shows that the small initial  $d_t$  value provides better results. The final uncertain optimized parameters are given in Table 4.2.2.

Table 4.2: Waveform and Distribution Design Under Uncertainty.

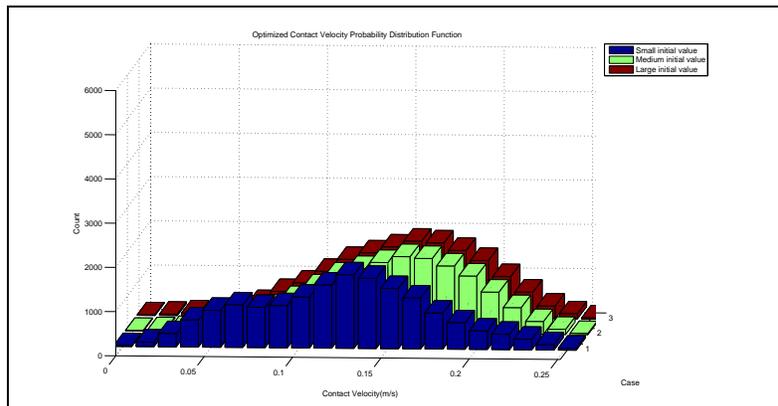
<b>Optimized Values:</b>	$V_h$	$t_b$	$E(d_t)$	$E(g)$	$E(V_0)$
Small Initial	40.34	8.13e-7	1.68e-7	1.81e-6	0.0
Medium Initial	45.7	7.78e-7	4.63e-7	1.81e-6	0.0
Large Initial	135.4	3.05e-7	1.20e-6	3.85e-6	0.0



(a) Uncertainty optimized closing times using contact velocity as objective function.



(b) Uncertainty optimized energies using contact velocity as objective function.



(c) Uncertainty optimized contact velocities using contact velocity as objective function.

Figure 4.3: Uncertainty optimized responses using contact velocity objective function.

### 4.3 Uncertainty Optimization over Normalized Energy

The final results to be discussed are from running our uncertainty optimization over our normalized energy metric,

$$J = \int_{t_{contact}}^{t_{close}} \dot{x}(t)^2 dt.$$

With the normalized energy as the optimization metric, the optimal waveform is one that closes the switch with no bounces. Thus this metric does not consider a higher impact velocity at contact to be a negative and may not be the best metric to use when considering device longevity. An example of the resulting displacement curve can be found in Figure 3.1.

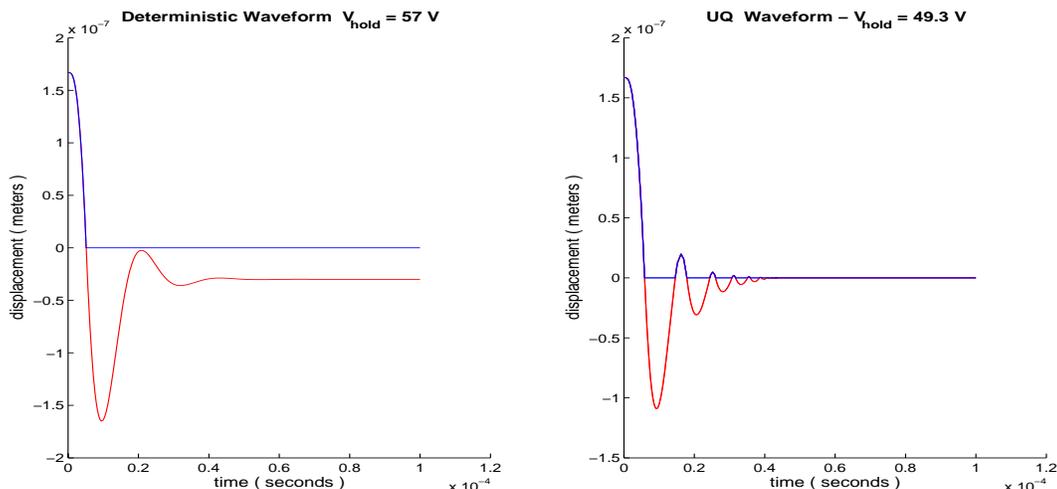


Figure 4.4: Comparing Displacement Curves of an Ideal Switch. Displacement curves from initial guess and optimized waveform for normalized energy optimization. The curve on the left has a holding voltage of 57V while the curve on the right has a holding voltage of 49.3V.

The initial guess for this run of the optimization procedure was given a travel distance of 1.668e-07 m, a gap distance of 1.73e-06 m, a holding voltage of 57 V and a build up time of 8.17e-07 seconds. This results in our normalized energy to be zero ( since we are integrating over a single point ) where a non-normalized energy would still consider the initial impact velocity.

The optimization procedure was allowed to vary both the holding voltage and the build up time while attempting to reduce the 90th percentile of the normalized energy under our uncertainty distributions. That is, the distributions themselves were not optimized in these results. The resulting optimum waveform has both a lower holding voltage ( 49.3 V compared to the initial guess of 57V ) and a longer build up time ( 8.5e-07 seconds compared to the initial guess of 8.17e-07 seconds ). While this "optimum" waveform has a resulting normalized energy that is practically the same as the initial guess ( the differences are in the range of round off error around 1e-14 ) the same as the initial guess waveform, the resulting waveform does reduce the 90th percentile of the contact velocity from 10.1 cm/s to 8.6 cm/s. A comparison of the displacement curves for the initial guess and optimized under uncertainty waveforms can be found in Figure 4.4. Also histograms comparing the spread of normalized energies and contact velocities from samplings given the initial guess waveform and the optimized waveform can be found in Figure 4.5.

## 5 Conclusion

In this report we discuss a one dimensional forced damped harmonic oscillator model governing the dynamics of an RF MEMS switch. We have fit model parameters from experimental data to account for contact dy-

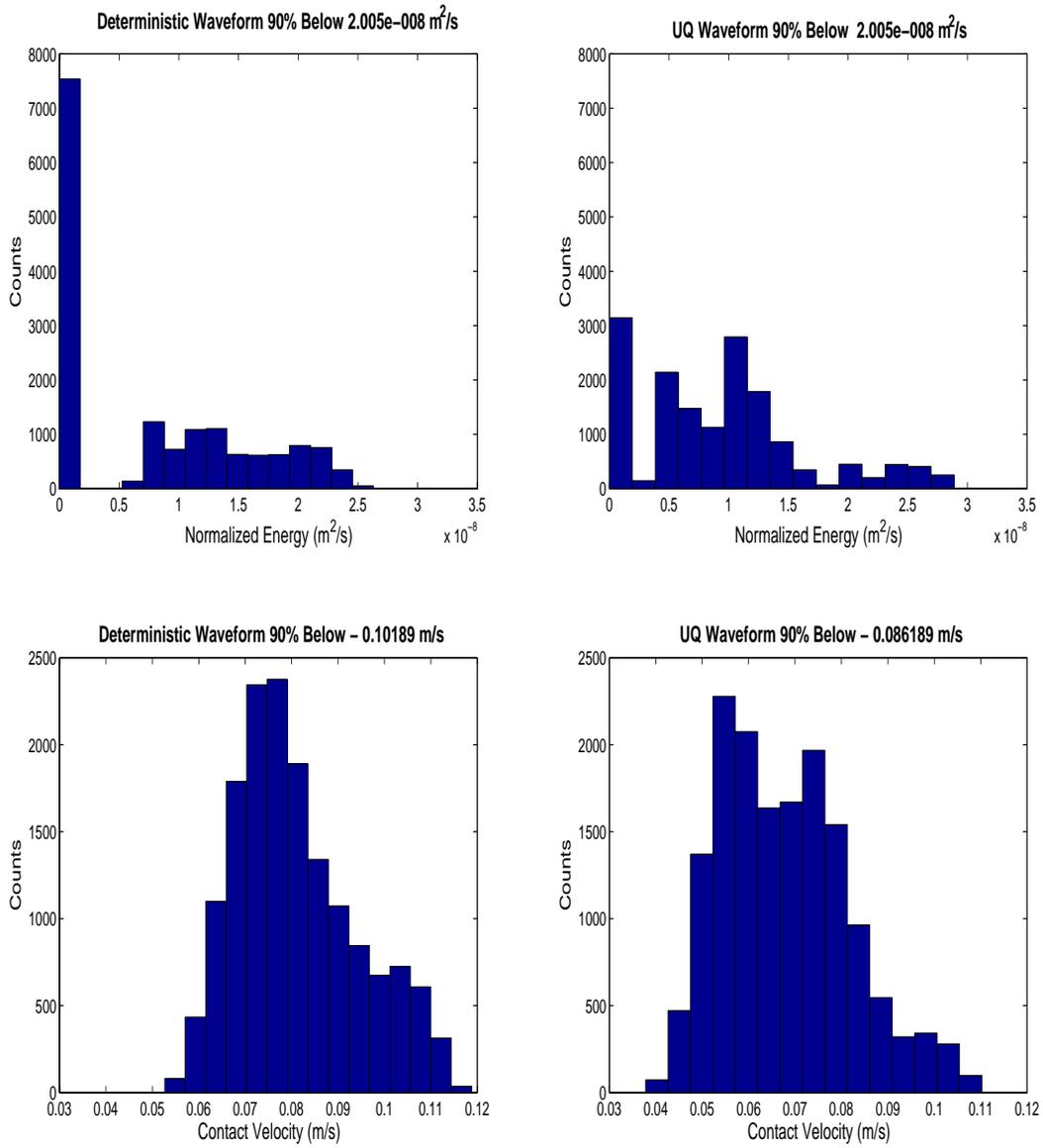


Figure 4.5: Comparing Samples of Waveforms with Uncertainty. Histograms of normalized energy and contact velocity samples from initial guess and UQ optimized waveform for normalized energy optimization.

namics. Through sensitivity analyses, we determine that gap, peak voltage and initial distance are sensitive model parameters that can significantly affect model output such as switch closing time and contact velocity. Parameters such as initial velocity and natural frequency are less sensitive. It is important to note that while model output is not sensitive to natural frequency, this parameter may play an important role in switch opening dynamics. Exploring this is left to future work. Once the sensitive model parameters are determined, we optimize the the voltage waveform over given distributions of the sensitive variables to perform an uncertainty analysis. This optimization returned a peak voltage and an RC time constant so the switch can be driven with a simple RC circuit voltage waveform. We have also found it is possible to design a MEMS switch that closes without rebounding.

Through the results obtained from this project we can make several recommendations to the manufacturers of RF MEMS switches. The first, is to use the smallest travel distance possible to make the switch with reliability. Smaller travel distances require smaller gap thicknesses and peak voltages to drive. We also recommend to use smaller RC time constants in conjunction with peak voltages, as these optimize model output. The next recommendation is to develop manufacturing processes that reduce variability in gap thickness and travel distance as these will produce more reliable switch dynamics and output. Lastly, consider the constraints of the dynamics associated with switch opening. The oscillatory behavior seen in switch opening could significantly affect switch closing performance.

## References

- [1] Allen, M., Massas, J., Field, R.J., Dyck, C.: Input and design optimization under uncertainty to minimize the impact velocity of an electrostatically actuated mems switch. *Journal of Vibration and Acoustics* **130** (2008)
- [2] Blecke, J., Epp, D., Sumali, H., Parker, G.: A simple learning control to eliminate rf-mems switch bounce. *Journal of Microelectromechanical Systems* **18**, 458–465 (2009)
- [3] Kaaajakari, V.: Mems tutorial: pull-in voltage in electrostatic microactuators. [Http://www.kaaajakari.net/ville/research/tutorials/pull\\_in\\_tutorial.pdf](http://www.kaaajakari.net/ville/research/tutorials/pull_in_tutorial.pdf)
- [4] Lucyszyn, S.: Review of radio frequency microelectromechanical systems technology. *Science, Measurement and Technology, IEE Proceedings* **151**(2), 93–103 (2004)
- [5] Nguyem, C.T.C., Katehi, L.P.B., Rebeiz, G.M.: Micromachined devices for wireless communications. *IEEE* **86**, 1756–1768 (1998)
- [6] Petersen, K.E.: Silicon as a mechanical material. *IEEE* **70**, 420–457 (1982)
- [7] Rebeiz, G.M.: *RF MEMS Theory, Design and Technology*. Wiley (2003)
- [8] Stanley, L.G., Stewart, D.L.: *Design sensitivity analysis: computational issues of sensitivity equation methods*. Society for Industrial and Applied Mathematics, Philadelphia (2002)
- [9] Sumali, H., Massad, J., Czuplewski, D., Dyck, C.: Waveform design for pulse-and-hold electrostatic acuation in mems. *Sensors and Actuators A* **134**, 213–220 (2007)
- [10] Yao, J.J.: RF MEMS from a device perspective. *Journal of Micromechanics and Microengineering* **10**, R9–R38 (2000)