

# How will we move: Modeling Climate-driven Age-specific Displacement Migration

Mathew E. Hauer \*

Department of Sociology, Florida State University  
and

Sunshine Jacobs

Department of Sociology, Florida State University

June 28, 2021

## Abstract

Global population projections show widespread ageing by century's end. The well-known relationship between age and migration propensity suggests more youthful populations are more likely to migrate than older populations. Yet relatively little research explores the important implication of ageing on human mobility in a changing climate, most likely driven by a paucity of migration data to comprehensively investigate the age structure of environmental migration. Here, we propose a displacement migration model to estimate age-specific migration probabilities. We combine a statistical outlier detection algorithm to first identify major environmental displacement events in US counties since 1980. Our findings suggests populations under age 60 are most likely to migrate after an environmental event, increasing with the size of displacement, validating many findings in the migration literature. We then build a flexible one-dimensional, age-specific displacement model based on these findings with estimation errors suggesting reasonable accuracy. Our proposed displacement model can be readily deployed to model prospective climate migration.

*Keywords:* Climate Change, Human Migration, Demography

---

\*Thanks y'all!

# 1 Introduction

## 2 Methods

First, we describe the data sets we used. Second, the methodology for detecting statistical outliers in time series. Finally, we describe the creation of our flexible one-dimensional, age-specific displacement model.

### 2.1 Data sets

We use two primary datasets: the National Vital Statistics System (NVSS) U.S. Census Populations with Bridged Race Categories data set<sup>1</sup> and the Spatial Hazard Events and Losses Database for the United States (SHELDUS).

The NVSS Bridged Race Categories data set harmonizes racial classifications across disparate time periods to allow population estimates to be sufficiently comparable across space and time. Importantly, all county boundaries are rectified to be geographically consistent across all time periods. We use the the 1969-2018 dataset which includes annual population estimates in five year age groups (0-4, ..., 85+), two sex groups (male and female), and three race groups (White, Black, Other).

In our statistical outlier analysis, we only consider counties created prior to 2000 and contained in the NVSS data. NVSS aggregated all counties in Hawaii to the state-level in the 1969-2018 NVSS bridged race data and we exclude them from our analysis. Several counties were created after 2000 (most notably is Broomfield County, Colorado). The 15 counties excluded from our analysis due to boundary changes or other reasons are Hoonah-Angoon Census Area AK 02105, Kusilvak Census Area AK 02158, Prince of Wales-Outer Ketchikan Census Area AK 02201, Skagway-Hoonah-Angoon Census Area AK 02232, Wrangell-Petersburg Census Area AK 02280, Adams County CO 08001, Boulder County CO 08013, Broomfield County CO 08014, Jefferson County CO 08059, Weld County CO 08123, Hawaii County HI 15001, Honolulu County HI 15003, Kalawao County HI 15005, Kauai County HI 15007, and Maui County HI 15009.

---

<sup>1</sup>Data can be downloaded here: <https://seer.cancer.gov/popdata/download.html>

We use these data in two separate steps. In our statistical outlier analysis, we aggregate all county-level estimates into annual total population estimates for each county for the period 1969-2016. The historical population estimates prior to 1980 display unusual volatility, so we consider only the time periods 1980-2018. We use the NVSS population estimates disaggregated by age/sex to calculate cohort-change ratios, which we describe below.

The second data source we use is SHELDUS [CITE]. SHELDUS is a county-level hazard data set for the US which contains information about the direct losses (property and crop losses, injuries, and fatalities) caused by a hazard event (thunderstorms, hurricanes, floods, wildfires, tornados, flash floods, earthquakes, etc.) for the period 1960 to the present. We use this database to ensure the county time periods we identify as statistical outliers with population losses contain experienced an environmental hazard in that county-year. This is to ensure the outlier population losses that we detect are associated with a hazard rather than other forces, such as economic forces.

## 2.2 Statistical Outlier Detection

We use a statistical time series outlier detection algorithm (?), implemented in the R programming language (?) via the tsoutliers package (?). This algorithm iteratively uses ARIMA models to 1) identify potential outliers or anomalies and 2) refit the ARIMA with the outliers removed to produce a counter-factual time series. Here we briefly summarize and describe the method.

Often, the behavior of a time series can be described and summarized in ARIMA models. If a series of values,  $y_t^*$ , is subject to  $m$  interventions or outliers at time points  $t_1, t_2, \dots, t_m$  with weights  $\omega$  then  $y_t^*$  can be defined as

$$y_t^* = \sum_{j=1}^m \omega_j L_j(B) I_t(t_j) + \frac{\theta(B)}{\phi(B)\alpha(B)} a_t \quad (1)$$

Where  $I_t(t_j)$  is an indicator variable with a value of 1 at observation  $t_j$  and where the  $j$ th outlier arises,

$\phi(B)$  is an autoregressive polynomial with all roots outside the unit circle,

$\theta(B)$  is a moving average polynomial with all roots outside the unit circle, and  $\alpha(B)$  is an autoregressive polynomial with all roots on the unit circle.

We examine three types of outliers at time point  $t_m$ :

1. additive outliers (AO), defined as  $L_j(B) = 1$ ;
2. level shift outliers (LS), defined as  $L_j(B) = 1/(1 - B)$ ; and
3. temporary change outliers (TC), defined as  $L_j(B) = 1/(1 - \delta B)$  where  $\delta$  is equal to 0.7.

Colloquially, additive outliers arise when a single event causes the time series to unexpectedly increase/decrease for a single time period; level shift outliers arise when an event causes the time series to unexpectedly increase/decrease for multiple time periods; and temporary change outliers arise when an event causes the time series to unexpectedly increase/decrease with lingering effects that decay over multiple time periods.

An outlier is detected with the estimated residuals using a regression equation

$$\pi(B)y_t^* \equiv \hat{e} = \sum_{j=1}^m \omega_j \pi(B) L_j(B) I_t(t_j) + a_t \quad (2)$$

where  $\pi(B) = \sum_{i=0}^{inf} \pi_i B^i$ .

Equations 1 and 2 allow for an automatic detection of anomalies iterated over a two-stage process.

In stage 1, anomalies are located. First, an ARIMA model is fit to the time series using the `forecast` package in R (??) where the best performing ARIMA model is selected based on the Bayesian information criterion (BIC). Next, the residuals from the forecast are checked for their significance using equation 2 where only anomalies above a critical  $t$ -static are considered “true” anomalies ( $|\tau| \geq 4$ ; p-value  $< 0.000063$ ). We chose this threshold to minimize the probability of committing a Type I error (or claiming an outlier is true when it is in fact not). Finally, two additional rules are implemented: If multiple anomalies are detected at the same time point, only the most significant anomaly is selected and if anomalies of the same type at consecutive time periods are detected, only the anomaly with highest  $t$ -statistic is selected.

In stage 2, anomalies are removed from the time series and a new ARIMA model is chosen and fit. The selection of the initial ARIMA model could have been affected by the presence of the anomalies, making some anomalies spuriously identified. To correct for this,

a new ARIMA model is fit accounting for additional regression effects in equation 1 from the list of candidate anomalies identified in stage 1, effectively removing the anomalies from the time series. Each anomaly is then reassessed under the new model and those anomalies that are no longer significant are removed.

These two stages are then iterated until no additional anomalies are detected.

**Figure 1** shows a toy example for anomaly detection in a time series using the example of Hurricane Katrina in Orleans Parish. On August 23 2005, Hurricane Katrina, a category 5 hurricane, struck southern Louisiana causing widespread damage and destruction in Orleans Parish in particular. The displacement from the Hurricane and the federal response were well documented (Hori et al. 2009, Fussell et al. 2014) and Census estimates suggest Orleans Parish lost more than 200,000 residents between 2005 and 2006. What would have been Orleans’ population estimate had Hurricane Katrina *not* occurred? A simple counter-factual estimate might keep population size just under 500,000 people ( $\hat{y}$ ).

In the Hurricane Katrina example, we have knowledge of the impact of Hurricane Katrina after the fact or ex-post-facto in order to create the counter-factual time series  $\hat{y}$ . But is this anomaly detectable without knowledge of the Hurricane Katrina? In other words, can we detect the reduction of Orleans Parish’s population between 2005 and 2006 using only the time series? Absolutely. The tsoutliers package identifies 2006 as an extremely strong level-shift outlier ( $t$ -stat = -96.39). The real world is hardly this simplified where a direct intervention is known, is large, and is testable.

## 2.3 Flexible one-dimensional, age-specific displacement model

We search all US counties for negative statistical outliers (indicating population losses) between 1980 and 2018. We detect population losses of magnitude  $4\sigma$  or greater. We then searched the SHELATUS database to see if these county-years experienced hazardous losses in excess of the 50th percentile. Four county-periods did not qualify. Additionally, one county-period contained age-sex groups with 0 people, necessitating exclusion.

With county-periods exhibiting large population declines after verified hazard losses, we build a flexible, one-dimensional, age-specific displacement model.

To link population displacement with age-specific population changes, we calculate

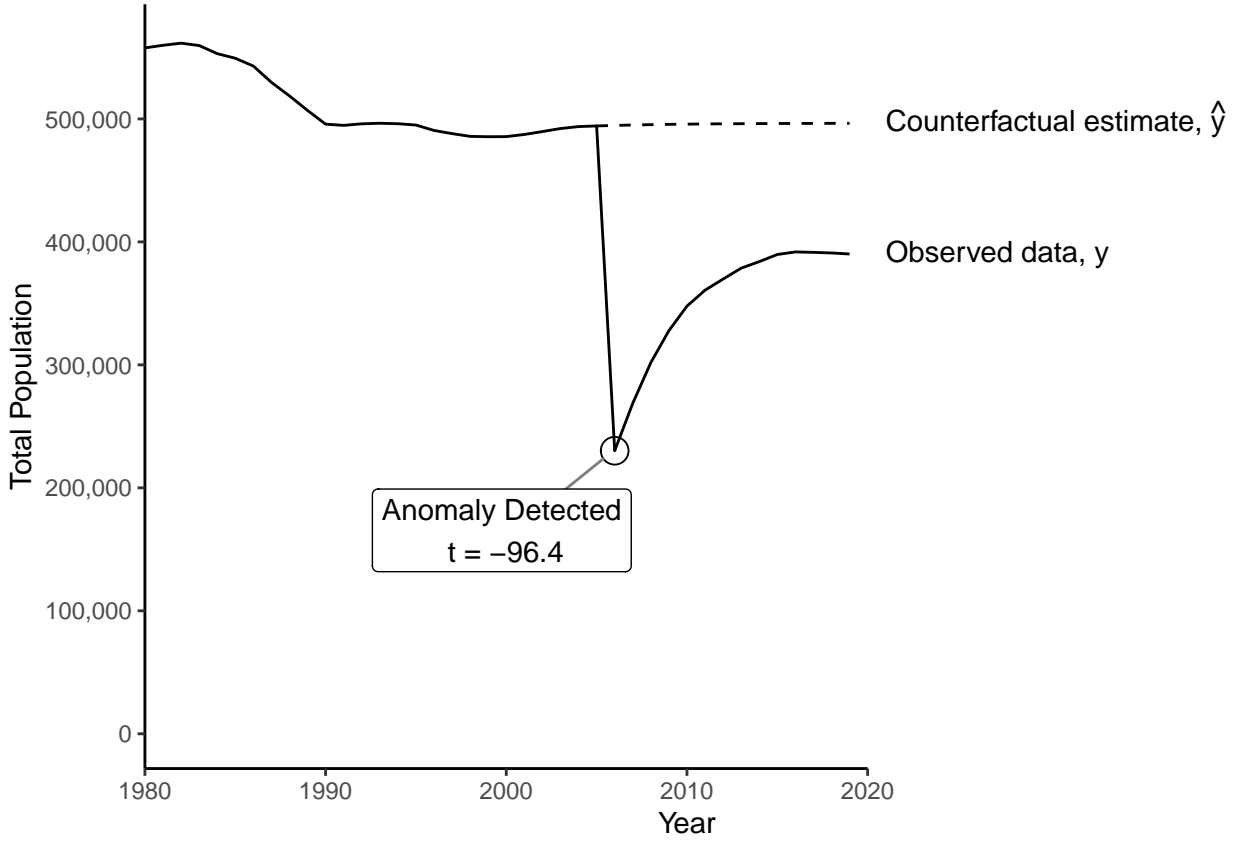


Figure 1: **An example of anomaly detection using Hurricanes Katrina and Rita in Orleans Parish Louisiana in 2005.** Hurricane Katrina struck Louisiana in 2005 and we use it as a toy example to illustrate our approach. This figure shows the annual time series of total population in Orleans Parish between 1980 and 2019. Between 1990 and 2005, Orleans Parish total population changed from 495k to 494k, suggesting a possible ‘plateau’ in the population (illustrated with the dotted ‘counterfactual’). Hurricane Katrina and the widespread population loss of more than 200k people represent a very strong anomaly ( $t=-96.39$ ).

cohort-change ratios in each county. Using the demographic accounting equation, the population at some time period  $t$  is equal to the  $P_{t-1} + B_{t-1} - D_{t-1} + M_{t-1,in} - M_{t-1,out}$ . For all age groups older than 0,  $B_{t-1}$  must be 0.

The calculation of a cohort-change ratio, as its name suggests, is relatively straightforward:

$$CCR_{x,t} = \frac{{}_nP_{x,t}}{{}_nP_{x-y,t-y}}$$

Where  ${}_nP_{x,t}$  is the population aged  $x$  to  $x + n$  in time  $t$  and  ${}_nP_{x-y,t}$  is the population aged  $x - y$  to  $x + n - y$  in time  $t$  where  $y$  refers to the time difference between time periods. Since mortality must decrement a population, any CCR above 1.0 implies a net-migration rate in excess of the mortality rate, and a growing population.

We build the following model based on the relationship between the change in CCRs at age  $x$ ,  $\Delta CCR_x = CCR_{x,t}/CCR_{x,t-1}$ , and the percentage decline in the total population compared to the counter-factual in the outlier detection method,  $\Delta P_t = \hat{P}_t/P_t$ :

$$\log(\Delta CCR_x) = a_x + b_x h + c_x h^2$$

Here,  $h$  is the  $\log(\Delta P_t)$  and shows a non-linear relationship with the logarithm of the change in CCRs by age.  $x$  refers to five-year age groups: 0-4, 5-9, ..., 85+.

### 3 Results

### 4 Conclusion

Table 1: **Correlation coefficients of changes in Cohort Change Ratios vs. Total Population Change (n=49).**

Age Group	Male	Female
0-4	0.931	0.944
5-9	0.954	0.934
10-14	0.963	0.965
15-19	0.806	0.930
20-24	0.650	0.919
25-4	0.710	0.934
30-34	0.752	0.959
35-39	0.874	0.975
40-44	0.936	0.964
45-49	0.971	0.963
50-54	0.969	0.977
55-59	0.965	0.958
60-64	0.896	0.888
65-69	0.809	0.912
70-74	0.937	0.826
75-79	0.839	0.795
80+	0.478	0.651



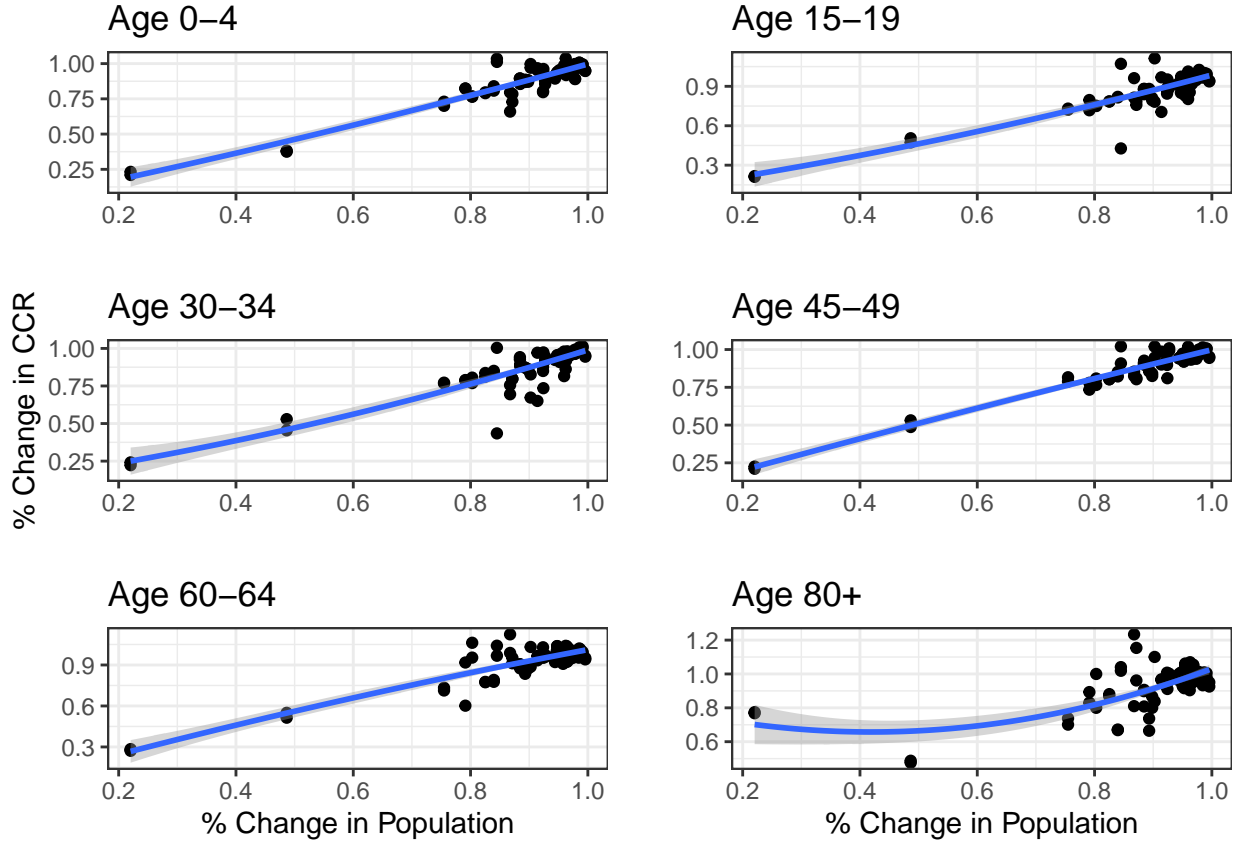


Figure 2: Age-specific changes in cohort-change ratios ( $\Delta CCR_x$ ) vs. change in total population ( $\Delta P_t$ ).

## References

- Fussell, E., Curtis, K. J. & DeWaard, J. (2014), ‘Recovery migration to the City of New Orleans after Hurricane Katrina: A migration systems approach’, *Population and Environment* **35**(3), 305–322.
- Hori, M., Schafer, M. J. & Bowman, D. J. (2009), ‘Displacement Dynamics in Southern Louisiana After Hurricanes Katrina and Rita’, *Population Research and Policy Review* **28**(1), 45–65.