Finding and Characterizing the Displaced: a method using administrative data

Jenna Nobles and Nathan Seltzer *

University of Wisconsin, Madison

In the aftermath of political, economic, or environmental disaster, documenting the welfare of affected populations serves goals in both policy and science arenas. Population displacement makes this task difficult, and sometimes impossible. We propose a method to document the welfare of the displaced that is inexpensive, quick-to-implement, and available whenever administrative data systems are minimally disrupted in areas neighboring sites of disaster or conflict. We use a series of Monte Carlo simulations to demonstrate when the method can be used. These simulations incorporate several varying dimensions, including disaster effect size, heterogeneity, and spillover; patterns of displacement; and features of the data available to the researcher. To further demonstrate the value of the approach, we apply the method to provide estimates of the impact of Hurricane Katrina on birth outcomes among displaced Gulf Coast residents.

*Center for Demography and Ecology, 1180 Observatory Drive, Madison, WI 53706. Abstract prepared for submission to the 2019 PAA meetings. The authors are grateful to Katherine Curtis, Eric Grodsky, Malia Jones, Jack DeWaard, and for computing resources provided by UW-Madison's SSC.

Background

Following large-scale environmental disaster or political conflict, relief and recovery efforts target the welfare of affected populations. Forced migration and population displacement make this task difficult. In some cases, the groups most in need of services are the hardest to find, particularly if they do not seek refugee or IDP services. A number of methodological tools have been developed for the study of populations that may not appear in traditional sampling frames or are difficult to locate for interview—sometimes referred to as "hidden populations" (Sudman, Sirkan, and Cowan 1988, Salganik and Heckathorn 2004, McCreesh, Frost, Seeley et al. 2012). These tools involve innovation in primary data collection, via ethnographic sampling and network leveraging for survey sampling. In the context of disaster specifically, new survey approaches also include large-scale efforts to locate and interview displaced members of affected communities sampled in pre-disaster data collection (Clark, Frankenberg, Sumantri, and Thomas 2014). These approaches are arguably the best practice for studying displacement and its effects on individual and community welfare. They are, however, expensive and time-intensive to implement.

In the present study, we propose a short-run, less expensive complement to these types of large-scale efforts. The proposed approach to locate and characterize displaced populations leverages existing administrative data systems for secondary data analysis. Administrative sources have several advantages. Displacing events are often difficult to predict and therefore require pre-existing data collection efforts to study. When displacement is localized in a small number of regions or when it not temporally aligned with rounds of national surveys, high-resolution enumeration of events or persons in administrative records facilitates population research. In addition, many administrative systems have high coverage rates (e.g., natality files cover over 99% of births in the United States), relative to forms of data collection that involve respondent refusal.

Of course, a key limitation of administrative data is that the information captured may be poorly aligned with the study of migration and displacement. Data on migration status or previous places of residence are rarely captured in records that have not been linked over time at the individual level. Nevertheless, we argue that systems of administrative data collection that are spatially and temporally referenced–i.e., that include data on time and place–can be used to detect displaced persons and approximate their characteristics.

The method we propose is most effective for characterizing the displaced following events

like environmental hazards or sociopolitical violence that are well defined in time and space. The method relies on data that describe counts of individuals (or uniquely-identifying data signatures of individuals, like a tax-record, a death certificate, or a flu shot) combined with the average traits of individuals by period (time) and region (place). Below, we describe the proposed method, a set of simulations to demonstrate the circumstances under which it may be most useful to researchers, and an empirical application in a contemporary population.

Approach

The proposed method proceeds in four steps: (1) We locate probable subpopulations of displaced persons by using discontinuities in time series count data on individuals in particular geographic regions. (2) We then estimate period-specific values of characteristics among the displaced (e.g., earnings, health) by leveraging the relative share of displaced persons per region and period against period change in the average values of those same characteristics observed for everyone in the region. Recall that the key identification challenge is that *individuals* cannot be identified as displaced. (3) We then use estimates of such characteristics with standard econometric tools to causally attribute *changes* in measures of wellbeing among the displaced to the event of interest (e.g., war, disaster). (4) Finally, we introduce a validation check. We use estimates of stable, time-invariant characteristics—traits that should not have been effected by the exposure (e.g., height, maternal education)—to validate that the displaced persons "found" by the method have values on stable traits that match their pre-displacement characteristics.

1. Finding displaced persons

When records are largely complete, we can locate displaced persons by identifying discontinuities in the time series of period- and region-specific counts of individuals. For example, consider counts of persons located in U.S. counties by month.

Formally, let N_{cm} = the number of persons in county c in month m.

Following a disaster event, N_{cm} for counties that contain displaced persons is composed of two groups, (1) persons who resided in that county prior to the disaster: N^r_{cm} and (2) disaster-displaced persons: N^d_{cm}. Of course, the resident population r includes a small change between month m-1 and m from montly net-migration that is not disaster-related. Therefore

N^r_{cm} is best understood as persons who are expected to reside in the county c in month m in the *absence* of the disaster.

We must learn the values of N^r_{cm} and N^d_{cm} when we only observe N_{cm}.

Conceptually, we seek an expected count of persons in affected counties had the disaster *not* happened. Positive deviation from this expected value may be interpreted as evidence of an influx of persons, under a few assumptions that we detail below. [1]

Identification of the month-to-month change in the expected count of persons in any county c must rely on two observations: past month-to-month trends in c which capture both the level and seasonal change in N^r_{cm}, as well as an estimate of the pre- post-disaster change in counts observed in counties that were not affected by the disaster. This latter estimate captures something distinct about the specific month-to-month change that spans the occurrence of a disaster but is not disaster-related. combined with period trends observed in largely unaffected, comparison counties.

To approximate N^r_{cm}, the analyst must distinguish plausible "recipient" counties–that is, places that displaced persons might plausibly have moved, from plausible "comparison counties–that is, places displaced persons likely did not arrive in that provide a plausible estimate of the counterfactual period trajectories in the outcomes of interest (more on this below). This is arguably the most subjective aspect of the method. This decision could be made using a number of types of information–e.g., proximity to the disaster-affected region, rapid response assessments of displacement, and so forth. There are, of course, implications of underestimating or overestimating the spatial spread of the region in which the displaced might be. If the analyst defines possible recipient categories too narrowly, the effect estimates may not capture the full population of displaced persons. If this is the case, non-random coverage error from this decision will appear in step 4 of the analysis, signaling error to the analyst. If the analyst defines possible recipient geography too broadly, the characteristics of the displaced will be estimated with less precision.

The analyst then pools count data for periods temporally spanning the disaster observed in the comparison counties and, one at a time, each of the receiving counties. This action creates n datasets, where n is the number of recipient counties. On each dataset, Eq. 1 is

[1] Of course, both positive and negative deviations will appear. Here, non-displacement deviations in both directions operate as statistical noise that reduce estimate precision.

estimated and returns δ_1 through δ_t, the deviation of the person count in recipient county c in the first month post-disaster through the t-th month post-disaster from that expected, given month-specific count trends in county c in the period prior to the disaster as well as any period change in counts observed in comparison counties.

$$\begin{aligned} N_{cm} = \alpha + \delta_1(\text{Month 1 post-disaster x Recipient}) ... + \delta_t(\text{Month t x Recipient}) \\ + \gamma_1(\text{Month 1 post-disaster}) ... + \gamma_t(\text{Month t post-disaster}) \\ + \beta_3 \text{Recipient} + \beta_4 \text{Year} + \beta_5(\text{Year x Recipient}) \\ + \text{Month fixed effects} + (\text{Recipient x Month fixed effects}) + \epsilon \end{aligned} \tag{1}$$

Anticipating step 2 of the method: the quantity needed to recover outcome estimates among the displaced is not the count of persons but the *proportion* of persons in each recipient county that is "unexpected" following the disaster. This is estimated using the total count of persons observed in the county and the values of δ_1 through δ_t in Eq. 1.

Expected persons for the county is equal to total persons minus unexpected persons, where $N^d_{c,\text{ Month 1 post-disaster}} = \delta_1$, Eq. 1:

$$N^r_{c,\text{ Month 1 post disaster}} = N_{cm} - N^d_{c,\text{ Month 1 post-disaster}} \tag{2}$$

The proportion of persons that is unexpected:

$$\pi^d_{cm} = \frac{N^d_{cm}}{N^d_{cm} + N^r_{cm}} \tag{3}$$

2. Recovering the outcomes of displaced persons

Let $\overline{X}_{cm}$ be the average value of an outcome measured in recipient county c and month m.

$\overline{X}_{cm}$ is a weighted average of two components: the average outcome value among expected residents, group r, X^r_{cm}, and the average outcome value among unexpected new residents of a county, group d, with outcome values X^d_{cm}:

$$\overline{X}_{cm} = X^d_{cm}\pi^d_{cm} + X^r_{cm}(1 - \pi^d_{cm}) \tag{4}$$

Rearranging:

$$X^d_{cm} = \frac{\overline{X}_{cm} - X^r_{cm}(1 - \pi^d_{cm})}{\pi^d_{cm}} \tag{5}$$

X^d_{cm} can be estimated with the observed values of outcomes in a given county, $\overline{X}_{cm}$, the proportion of persons that are unexpected given past trends, π^d_{cm}, and outcomes among expected persons X^r_{cm}. $\overline{X}_{cm}$ is observed in the data and π^d_{cm} is given in Eq. 3 in the section above.

Conditional on the assumption that the disaster did not negatively or positively impact non-displaced persons (an assumption we will later relax), then the expected value of X^r_{cm} might reasonably be its predicted value based on observed values leading up through the disaster month, combined with any period change observed for the country as a whole.

The analyst estimates X^r_{cm} in a similar manner to N^r_{cm}. That is, X^r_{cm} is the expected value of X_{cm}, had the disaster not occurred. This value can be estimated by using past trends within the county and observed period trends in comparison counties. Specifically, X^r_{cm} is the predicted value of θ^*_{cm} where the δ^*_n terms are not used in the prediction:

$$\begin{aligned}\theta^*_{cm} = \alpha^* + \delta^*_1(\text{Month 1 post-disaster x Recipient}) ... + \delta^*_t(\text{Month t post-disaster x Recipient}) \\ + \gamma^*_1(\text{Month 1 post-disaster}) ... + \gamma^*_8(\text{Month t post-disaster}) \\ + \beta^*_3\text{Recipient} + \beta^*_4\text{Year} + \beta^*_5(\text{Year x Recipient}) \\ + \text{Month fixed effects} + (\text{Recipient x Month fixed effects}) + \epsilon^*\end{aligned} \tag{6}$$

With values of X_{cm} for each receiving county in hand, it is also possible to recover the average value of any given outcome for *all exposed persons* distributed across recipient counties, X^d_m, with Eq. 7, where $\rho_{cm} = \frac{N^d_{cm}}{\sum_{n=1}^{c} N^d_{cm}}$, the proportion of all displaced persons located in recipient county c.

$$\theta_m = \sum_{n=1}^{c} \rho_{cm} X^d_{cm} \tag{7}$$

Eq. 7 is a weighted average of estimates of outcomes to displaced persons found in recipient county c and month m, where the weights are the *proportion* of all displaced persons found in recipient county c and month m. Note that when some persons in a disaster-affected region are not displaced, Eq. 7 will also include information observed on persons still residing in the sending county as one of the counties used to calculate θ_m.

3. Recovering the effect of the disaster on outcomes among those exposed

With the post-disaster characteristics of the displaced population and the total exposed population in hand, it is also possible to test whether the recovered average value for the affected population differs from the value that would have been expected had the disaster not occurred–that is, what the effect of the disaster is on the welfare of the exposed population. To do this, the analyst combines the values of θ_m generated in Eq. 7, which are the post-disaster observations of exposed persons, with the pre-disaster values of exposed persons observed in the data as well as the pre- and post-disaster values of persons in comparison counties observed in the data.

A standard approach to estimating disaster effects is to use a difference-in-difference estimator, which relies on pre- and post-disaster observations of the population of interest ("exposed") and an untreated ("comparison") population. The comparison population provides an estimate of period change that might have been expected had the disaster had not happened. In a regression framework, this estimate can be derived from an equation like the following, estimated on a set of pooled observations of exposed and unexposed counties:

$$\begin{aligned} \theta_{cm} = \alpha + \beta_1(\text{Post x Exposed}) + \beta_2\text{Post} + \beta_3\text{Exposed} \\ +\beta_4\text{Year} + \beta_5(\text{Year x Exposed}) \\ +\text{Month fixed effects} + (\text{Exposed x Month fixed effects}) + \epsilon \end{aligned} \tag{8}$$

Here β_1 provides the estimate of interest: the difference between the observed and expected values in exposed regions.

4. Validating the approach

The approach described here is designed to recover information about displaced persons, even when they are not directly identified as displaced persons. One way to validate this approach is by testing whether it is possible to recover information about displaced persons that *should not be affected by the exposure of interest.* Such characteristics might include completed education among older adults, parental age at birth, or even stable health characteristics, like height. Validating the approach involves using the series of eight equations described above, but replacing the time-varying outcome measures of individual welfare with stable measures of mothers' education or age. Conceptually, this exercise asks if it is possible to

reconstruct the expected value of stable traits. In this case, a precisely estimated β_1 of zero indicates recovery of these characteristics.

When is this method effective?

Displacement events vary on many dimensions, as do the data systems available to study them. To demonstrate the circumstances during which this method is effective at recovering characteristics of the displaced, we use a series of Monte Carlo simulations in which we subject a hypothetical population to a perturbation with effect size β and assess whether, and with what confidence we can recover β. We allow several types of features to vary in these simulations, including: (a) characteristics of administrative data sources, like size, number of units, and temporal resolution; (b) patterns of displacement, including degree of spatial diaspora clustering, and (c) characteristics of exposures, including presence of spillover effects and effect heterogeneity. We detail this process below; the seven relevant parameters are summarized in Table 1.

Table 1: Parameters used in Monte Carlo simulations

	Parameter	Values	Representing
1	Geographic Unit	50, 500, 1,000	Provinces, States, Counties
2	Time periods	5, 20, 50	Years, Quarters, Months
3	Sampled observations	50 - 5,000	ACS, Vital Statistics, etc
4	Spatial clustering of displaced	Even - Clustered	
5	Effect size	$\beta = 0.1, 0.25, 0.5, 1.0$ SD of X	
6	Effect heterogeneity	Variance = 0 or $\beta/4$	
7	Spillover effects on non-displaced	$\beta = 0$ or $\beta/4$	
8	Proportion of residents displaced	Distribution of π^d_{cm}, Eq. 3	Calculated from (1) - (4)

The simulations are initiated with a draw of a hypothetical population situated across c=50, 500, or 1,000 geographic units–these could represent provinces, states, counties, municipalities, neighborhoods and so forth. We generate sample data from these units for m =5, 20, or 50 consecutive time periods. Within each unit and each time period, we observe a sample of n =50, 100, 500, 1000, or 5000 persons. These persons have a value on a characteristic of interest, X, that is normally distributed within the geographic unit and shifted with a random component, ϵ_c that varies across units but is constant across time periods within each unit,

and a random component that varies across units and time periods, ζ_{cm}. ϵ_c is intended to represent stable mean differences in X across geographic units–for example, that average BMI values are higher in Alabama than in Colorado or that average earnings are higher in San Francisco than in Des Moines. ζ_{cm} captures a stochastic process generating noise in the data.

The combinations of variability in the number of units, the periods of observation, and the number of persons observed in each unit are meant to capture plausible ranges of sample size and density observed in currently-available data sources.
The simulations continue with a disaster of some type. In the m+1 period, the residents of one geographic unit experience an event that (i) causes displacement and (ii) shifts the characteristic of interest X by a factor of β.

Here we consider two types of variation. When the majority of persons displaced from one region cluster in another location, inferential reconstruction will almost certainly be more precise relative to a scenario when displaced persons are distributed across receiving locations–which effectively reduces the ratio of displaced persons to residents (π^d_{cm} from Eq. 3, above). We thus consider three different degrees of spatial clustering. In each case, we assume that at least one displaced person is located in half of the geographic units sampled. The clustering scenarios are captured in Fig 1, and range from a distribution in which 30% of residents arrive in one single other geographic unit, to an equal distribution of residents across receiving units. Note that the effect of variation on clustering parameter interacts with the sample size and the number of relevant geographic units to shift π^d_{cm}.

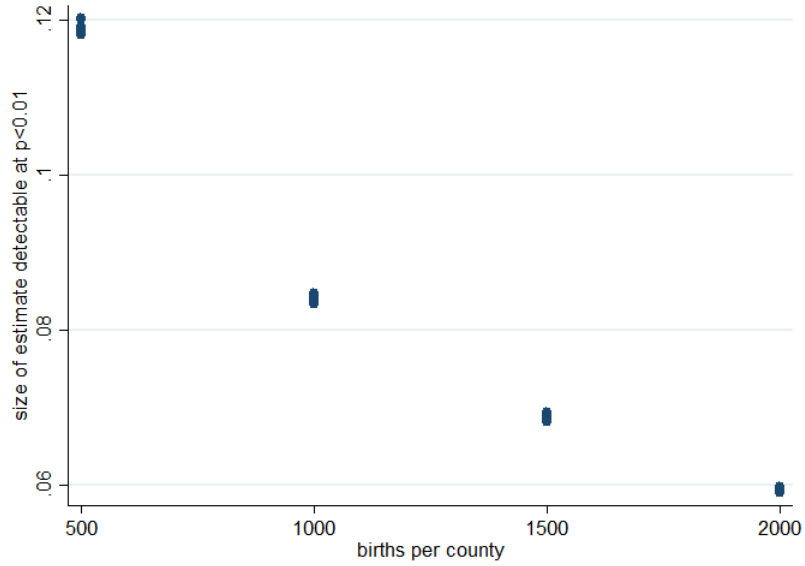
Finally, we consider two types of variation in the disaster's effect on the population characteristic X. We test cases in which the effect β is experienced equally by all displaced persons, and cases in which β varies at the individual level. In this latter scenario, β_{icm} for persons in the exposed region is a draw of a normally distributed variable with mean β and variance $\beta/4$. We also consider that some economic, political, or environmental disasters have spillover effects on non-displaced persons outside of the "exposed" geographic region. We consider scenarios in which β_{icm} is non-zero for persons in one-quarter of the regions receiving displaced persons. These spillover effects are one-tenth and one-quarter as large as β_{icm} for persons in the exposed region.

In sum, then, the simulations incorporate 7 parameters: the (1) number of geographic units, (2) number of periods of observation, (3) number of persons observed per region-period combination, (4) spatial clustering of displaced persons, (5) disaster effect size, (6) disaster effect

homogeneity, (7) spillover of the disaster effect onto non-displaced persons. We keep track of–and report–a 8th parameter, π^d_{cm} that is generated by the interaction of (1), (2), (3), and (4) because this may be the most relevant for other analysts considering the method. The resulting possible combinations presents a broad, flexible parameter space designed to capture a wide range of situations combining types of data and types of disasters. For each of the 3240 combinations of parameters 1-7, we draw 1000 simulations and generate a distribution of β returned by the simulation.

As an example, Figure 1 below displays the smallest magnitude of exposure effect (β_1 in Eq. 8 in standard deviations) that can be detected at p<0.01 across the number of administrative records per geographic region and per unit of time. In the aforementioned example, this is a range of the average births per county per month. Figure 1 displays a small fraction of the parameter space these simulations explore.

Figure 1: Magnitude of Exposure Effect Detectable at p≤0.01 with Proposed Method by Number of Observations per Time-Region Classification



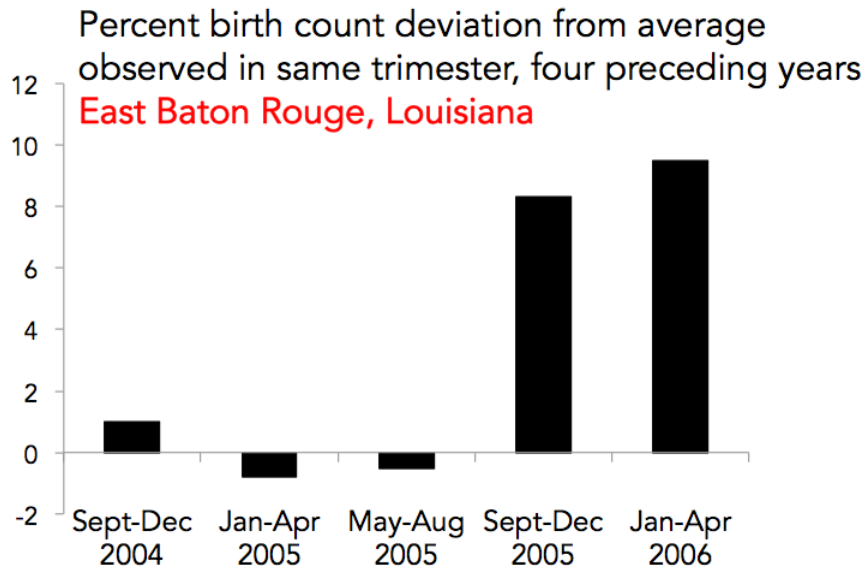
Application: Pregnancy and Infant Health Outcomes in Displaced Gulf Coast Families

In ongoing research, we illustrate the approach by describing a particular application: pregnancy and infant health outcomes among births to women displaced by Hurricane Katrina.

Katrina landed on shores of the Gulf Coast as a category 3 hurricane on August 29, 2005. Thirteen hundred people were killed, 108 billion dollars of property was damaged. Approximately half a million people were displaced. The major birthing centers in New Orleans closed. In September 2005, 2 births occurred in New Orleans, both at home. Instead pregnant women gave birth in new locations (Figure 2).

A number of scholars have devoted years to studying the spatial dispersion of Gulf Coast residents as a result of the disaster (DeWaard et al. 2016, Fussell et al. 2010, Groen and Polivka 2010, Sastry 2009, Sastry et al 2014). This choice of application facilitates the comparison of estimates from the proposed method with those from other scholarship–including research using one-year migration estimates from the ACS and the spatial patterning of FEMA applications.

Figure 2:



Note: authors' calculations, NCHS vital statistics data.

The new empirical information provided by applying the method to the case of Katrina-affected Gulf Coast residents is thus not the patterns of displacement but rather the reconstruction of characteristics among the displaced that are not available in other data sources—like pre-term delivery, delivery complications, and infant health—outcomes that are potentially responsive to maternal experiences of disaster and displacement.

References

DeWaard, J., Curtis, K. J., and Fussell, E. 2016. Population recovery in New Orleans after Hurricane Katrina: exploring the potential role of stage migration in migration systems. Population and Environment, 37(4), 449463.

Fussell, E., Sastry, N., and VanLandingham, M. 2010. Race, socioeconomic status, and return migration to New Orleans after Hurricane Katrina. Population and Environment, 31(1-3), 2042.

Gray, Clark, Elizabeth Frankenberg, Cecep Sumantri, and Duncan Thomas. 2014. Studying Displacement after a Disaster Using Large Scale Survey Methods: Sumatra after the 2004 Tsunami. Annals of the Association of American Geographers.104(3): 594?612.

Groen, J. A., and Polivka, A. E. 2010. Going home after Hurricane Katrina: Determinants of return migration and changes in affected areas. Demography, 47(4), 821844.

McCreesh, N., Frost, S., Seeley, J., et. al. 2012. Evaluation of respondent-driven sampling. Epidemiology, 23(1), 138-47.

Salganik, M. J. and Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. Sociological Methodology, 34, 193-239.
Sastry, N. 2009. Displaced New Orleans Residents in the Aftermath of Hurricane Katrina: Results from a Pilot Survey. Organization & Environment, 22(4), 395409.

Sastry, N., Gregory, J., Sastry, N., and Gregory, J. 2014. The Location of Displaced New Orleans Residents in the Year After Hurricane Katrina. Demography, 51, 753775.

Sudman, S., Sirken, M., Cowan, C. D. 1988. Sampling rare and elusive populations. Science, 240(4855), 991-6.