Mining Administrative Data to Spur Urban Revitalization*

Ben Green
Harvard University
Cambridge, MA
bgreen@g.harvard.edu

Alejandra Caro
Carnegie Mellon University
Pittsburgh, PA
alejandra.caro@moodys.com

Matthew Conway†
University of Chicago
Chicago, IL
matt@indicatrix.org

Robert Manduca
Harvard University
Cambridge, MA
rmanduca@g.harvard.edu

Tom Plagge
University of Chicago
Chicago, IL
tplagge@gmail.com

Abby Miller
Innovation Delivery Team
Memphis, TN
abby.miller@memphistn.gov

ABSTRACT

After decades of urban investment dominated by sprawl and outward growth, municipal governments in the United States are responsible for the upkeep of urban neighborhoods that have not received sufficient resources or maintenance in many years. One of city governments' biggest challenges is to revitalize decaying neighborhoods given only limited resources. In this paper, we apply data science techniques to administrative data to help the City of Memphis, Tennessee improve distressed neighborhoods. We develop new methods to efficiently identify homes in need of rehabilitation and to predict the impacts of potential investments on neighborhoods. Our analyses allow Memphis to design neighborhood-improvement strategies that generate greater impacts on communities. Since our work uses data that most US cities already collect, our models and methods are highly portable and inexpensive to implement. We also discuss the challenges we encountered while analyzing government data and deploying our tools, and highlight important steps to improve future data-driven efforts in urban policy.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Statistical computing; H.2.8 [**Database Management**]: Database Applications—*Data mining, Spatial databases and GIS*; J.1 [**Administrative Data Processing**]: Government

Keywords

Neighborhood distress; Social good; Urban revitalization

*This work was completed as part of the 2014 Eric and Wendy Schmidt Data Science for Social Good Summer Fellowship at the University of Chicago, in partnership with the City of Memphis, Tennessee.

†Previous affiliation.

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

ACM 978-1-4503-3664-2/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2783258.2788568.

1. INTRODUCTION

US cities face a legacy of rapid outward growth and suburbanization. Since the middle of the twentieth century, municipalities across the country implemented policies designed to increase their land area, population, and tax base. As cities invested in growth, however, they neglected to properly maintain their older, central neighborhoods. Populations declined and the housing stock deteriorated, leaving struggling urban cores throughout the United States.

Recent efforts by governments and communities to reinvest in central city neighborhoods have been uneven, especially since the financial meltdown of 2008. While some neighborhoods have rapidly revitalized, many others continue to struggle with foreclosures, crime, and joblessness. These neighborhoods often require more money to deliver services than they generate back in fees and tax revenues, draining city resources. Revitalizing distressed neighborhoods is therefore a primary goal in many American cities.

We define distress as properties in need of structural or cosmetic repairs in order to be brought up to the community's standards. Although distress is perhaps more commonly referred to as blight, we avoid that term here given the negative connotations associated with it. We use distress to refer simply to the physical condition of homes, not as a political label assigned to neighborhoods.

Distress is a symptom of economic malaise, but potentially a cause as well. Distressed properties themselves produce little or no tax revenue and send negative signals to communities and investors. They can depress the value of nearby homes and diminish their neighborhood's quality-of-life [16, 20]. Distressed and abandoned properties also limit social interactions among neighbors, a result linked to higher crime rates, decreased public health, and other negative indicators of community well-being [18].

However, redeveloping traditional core neighborhoods is a challenging goal. The lots and homes tend to be relatively small, many areas face long-standing issues with poverty and crime, and some properties come with tax liens that raise the effective cost to purchase the property. As a result, it often takes encouragement from public or nonprofit investors to kickstart private development in these neighborhoods.[1]

[1]While policymakers must be mindful to avoid causing gentrification or displacement, the more pressing concern in most severely-distressed neighborhoods is improving conditions up to a livable standard.

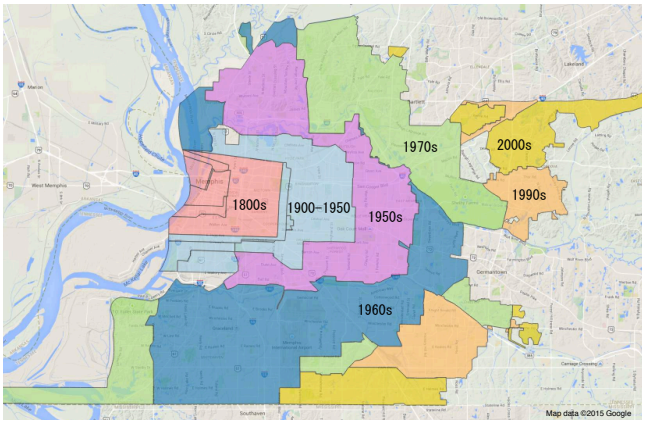

Figure 1: The land area of Memphis has grown more than 350% since 1950 due to annexations.

The City of Memphis, TN is a case study for many of these issues. Over the past several decades, Memphis has pursued a growth strategy of annexing land for suburban-style greenfield development. From 1950–2010, the City's geographic area increased 350% (Figure 1). Yet, in that same timespan, the City's population grew only 161% [3]. Furthermore, Memphis contains more of its regional population than many other American cities (approximately 50%, compared with 20% in Boston), meaning that it is responsible for maintaining a larger portion of its sprawl [1]. This has left the city with a hollowing core and sprawling population. The City's resources are, quite literally, stretched too thin across its geography: a recent report declared "Memphis literally does not have the financial resources to continue with this pattern of growth" [13].

To combat this trend, a coalition of groups including Memphis government agencies and local Community Development Corporations (CDCs) have begun investing in portions of the city's traditional urban core. From the City's perspective, rehabilitation and construction in core neighborhoods are highly productive: they make use of existing infrastructure, increase property tax revenues, and generate additional income from sales taxes and fees. For this and other reasons, the city would like to see such renewal continue.

Yet given the limited resources that governments and CDCs have at their disposal, they must be selective about where to invest money. They want to find the properties and determine the actions that will most impact property values and social well-being.

The work described in this paper was completed in order to aid these efforts in Memphis. We worked with the Innovation Delivery Team, a group in the Mayor's Office funded by Bloomberg Philanthropies that brings innovative approaches to pressing and complex urban challenges.[2] We gathered data from various government departments and neighborhood organizations, which we then analyzed to develop data-driven tools the City can use to improve its neighborhood revitalization strategies.

Of course, urban development cannot simply be optimized using data and algorithms. Politics, funding, community engagement, and individual initiative all impact what actions are possible and expedient. Yet data-driven strategies can help communities better understand which decisions will yield the most beneficial impacts. Proposing policies that are backed by good data and analysis offers useful insights that can inform stakeholders throughout the planning process and increase the likelihood of implementation.

We focus here on three aspects of the problem: identifying distressed properties, characterizing their effects on neighborhoods, and assessing the costs and benefits of remediation strategies. We first describe previous related work and the data used for our analyses (Sections 2 & 3). Section 4 describes the tools we developed that mine neighborhood surveys and administrative data to identify distressed homes. In Section 5, we discuss our efforts to evaluate the impacts of distressed homes and different investment strategies on neighborhood property values. After highlighting our primary impacts in Memphis in Section 6, we discuss in Section 7 important lessons learned while conducting data science analyses for municipal governments. Finally, we discuss future work in Section 8 and conclude in Section 9.

All of our code and some data used in our analyses are available at `https://github.com/dssg/memphis-public`.

2. RELATED WORK

Many studies have attempted to quantify the effects of distressed and abandoned homes on neighborhood property values. A 2001 study conducted in Philadelphia found that "all else being equal, houses on blocks with abandonment sold for $6,715 less than houses on blocks with no abandonment" [16]. The same study also warns against demolishing homes, finding that rehabilitating properties was more effective at stabilizing neighborhoods.

Analyses of vacant homes find that these properties have far-reaching detrimental effects on cities: they are threats to public safety, drive down neighborhood property values, and decrease the community's quality-of-life [20]. Among the challenges identified for cities dealing with this issue is a lack of data about abandoned properties and their costs.

One technique used to target investments within cities is Market Value Analysis. Looking at a variety of neighborhood indicators, cities can use clustering algorithms to segment all of its neighborhoods into a small number of categories (these might include "strong and growing" and "weak and declining"). Cities can then develop common strategies to implement in multiple neighborhoods of the same category. Previous efforts using this approach have been successful in Philadelphia, Baltimore, and other cities [9]. Another study of neighborhood improvement strategies found that revitalization efforts are most successful when investments are clustered in a few neighborhoods rather than distributed evenly across the city [7].

Within Memphis, many partners are working to revitalize core neighborhoods and combat distress. Organizations such as Livable Memphis, Community LIFT, and The Economic Development Growth Engine (EDGE) have spent the past several years implementing programs that emphasize growth, sustainable communities, and economic development.

Starting in October 2012, the Innovation Delivery Team began developing a series of initiatives that revitalize neighborhoods through a simple formula: "clean it, activate it, sustain it." These efforts work in three stages:

Clean Eliminate physical barriers to investment.

[2] `http://innovatememphis.com`

Activate Deploy small-scale and temporary changes to inject energy and demonstrate what is possible.

Sustain Make successful changes permanent through public policy.

To test various approaches to neighborhood vitality, the Innovation Delivery Team has so far focused their efforts on three neighborhoods: South Memphis, Binghampton, and Crosstown. Thus far, the Innovation Delivery Team's work shows the City of Memphis how moderate investments applied in focused ways can generate significant returns for neighborhoods.

3. DATA DESCRIPTION

3.1 Administrative Data

Most of the data we use in this paper was collected by local administrative bodies that regularly gather information about property conditions and neighborhood well-being. One important factor to keep in mind is that this data was not originally collected for assessing and targeting urban revitalization, making the incremental data collection cost for this type of work negligible. While collecting additional data would improve the models we describe later, it would be quite burdensome. By using data that already exists in almost every US city, we highlight methods that are possible across the country without requiring extensive data-collection efforts.

The majority of our data came from the City's internal database of every parcel (plot of land) in Memphis. Our primary source of information was property assessments, which contains information such as appraised value, parcel size, number of rooms, and building condition. This allowed us to consider many specific aspects of each property that may help us understand its health.

Another valuable dataset from the City of Memphis details which properties have had their utilities disconnected, been subject to code violations, or fallen into tax delinquency. These all indicate some level of poor maintenance and so are all valuable for identifying properties in need of reinvestment.

We also obtained data on every foreclosure that occurred in Memphis since 2000. Foreclosures are a clear indicator of distressed properties and financially-struggling homeowners. Because this data came from a separate source than our other property data, we were unable to reliably match foreclosures with specific properties. Instead, we were only able to aggregate foreclosures based on distance to properties and Census block groups.

Our final administrative data sources were the US Census and American Community Survey. Neighborhood-level statistics are valuable for understanding the context of how a given property fits into the fabric of the city.

3.2 Neighborhood Surveys

Critical to our work were several neighborhood windshield surveys conducted by community groups in Memphis. Most useful was the Neighborhood-by-Neighbor survey, which was conducted by The Center for Community Building and Neighborhood Action (CBANA) at the University of Memphis between February 2008 and January 2010 [5]. Volunteers familiar with each neighborhood surveyed the entirety of Memphis, identifying all residential properties not in compliance with the City's anti-blight housing code. Similar surveys were conducted by the Binghampton and Frayser Neighborhood CDCs in 2011 and 2013, respectively.

Although in-person inspections are the most reliable method for labeling property conditions, there are nonetheless a few caveats when using data of this kind. Surveys are completed by volunteers rather than trained professionals, and each person only surveys a small portion of the city. While the organizations running neighborhood surveys typically train volunteers and provide common definitions, there remains potential for inconsistent classifications. This is especially true when comparing results across different surveys.

4. IDENTIFYING DISTRESSED PROPERTIES

A major challenge for cities dealing with distressed neighborhoods is determining which properties need repair. While the identities of a city's most struggling neighborhoods are often widely-known, there is no simple way to determine the condition of specific properties without in-person inspections. Revitalizing communities typically requires a small number of targeted investments, however. If city officials had better data about property conditions, they would be able to invest more effectively.

The most common approach to gather property condition data is a windshield survey, in which a government department or community group organizes volunteers to visually inspect the outside of each home.[3] Windshield surveys are time-consuming and expensive, however, and often require countless volunteer hours and private funding from donors. It is therefore impractical for cities to rely on such surveys to diagnose local housing conditions.

In order to alleviate the need for regular surveys, we designed a system that uses administrative data to estimate the risk that each residential property in Memphis is distressed. We assembled pre-existing data, collected by the City of Memphis for other purposes such as property assessments and utility payments, and built a model by comparing these data to survey results.

While such a model is not as accurate as in-person inspections, relying exclusively on administrative data has an important benefit: it takes little time or money to derive new estimates. All of the data used as features in our model are already collected regularly by the City for other purposes. Estimates of property conditions can therefore be updated annually with little overhead, allowing the City and other stakeholders to efficiently track trends.

This tool obviates the necessity to conduct resource-intensive full-city surveys on a regular basis. Use of the model will highlight hotspots and outliers to help the City set priorities for thorough in-person surveys without requiring expensive preliminary work.

4.1 Data and Features

We built our model using thirty input features contained in the data described in Section 3. These include:

- Total appraised value of the home
- Age of the home

[3]This has traditionally been done from within a car, hence the term windshield.

- Whether the home has had its utilities disconnected in the past
- Percent change in the assessed value of the home over the past four years
- Percent of properties in the Census block of the home that were foreclosed on in the past year

We trained our model using the Neighborhood-by-Neighbor windshield survey conducted in Memphis [5]. This data served as our ground truth and allowed us to identify if a property was distressed or not in 2008.

4.2 Model

We trained a random forest classifier to predict, for multiple years, whether each residential property in Memphis was distressed. The model was built using the `randomForest` [11] package in `R` [15]. We also experimented with other classifiers, specifically decision trees and logistic regression, but we don't report the results here since random forests produced the best results.

Because the labeled data we have for the entire city came from the Neighborhood-by-Neighbor survey in 2008, we used that year's data as our training set. While the survey categorized distressed properties into seven different categories, we reduced the labels to a binary classification of "distressed" and "not distressed." Due to the relatively small sample size for each distressed class and the potential for volunteer surveyors to label similar properties inconsistently, we did not feel confident we could train the model to accurately distinguish between multiple forms of distress. Future work can involve selectively building more detailed models if necessary to classify properties into more groups.

Nonetheless, our estimates of a property's condition were more nuanced than binary classification because we considered the predicted class probabilities rather than just the predicted class. In a random forest classifier, each property in the test set is compared with those most similar in the training set. Roughly speaking, the predicted class probabilities reflect the proportion of similar properties from the training set labeled with each class. If a property is labeled as distressed with a probability of 0.75, for example, 75% of the most similar properties in the training set were labeled as distressed while the other 25% were labeled as not distressed. We define a property's class probability for being distressed as its risk score.

4.3 Validating the Classifier

In addition to testing the classifier using cross-validation (described below), we first wanted to get a more qualitative sense of what our predictions reveal about actual homes. We shared the results of the model with our collaborators in Memphis to get their feedback on some of the properties we had classified.

We then "walked" a few neighborhoods on Google Street View to compare the images for a random sample of properties with the risk scores calculated by our model. The images viewed were taken in August 2013. On one block in the Midtown neighborhood, for example, risk scores range from 0.269 to 0.807. Figure 2 shows two example homes with high and low risk scores. Clear differences are visible between these two properties, such as the differently maintained lawns.

Figure 2: There is a visible difference between homes with a low risk score (0.368, left) and a high risk score (0.807, right).

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
AUC	0.829	0.832	0.829	0.832	0.832

Table 1: The results of cross-validation were consistently accurate.

While this test clearly does not validate our model, observing many homes in this manner encourages us to be confident in the model's predictive ability. In general, we find that our risk scores provide accurate insights into the relative conditions of properties.

4.3.1 Cross-Validation

We validated our model using five-fold cross-validation. We randomly selected the data into five groups, and then trained a model on four-fifths of the data and tested on the last fifth. The accuracy of our model was evaluated using a precision-recall curve [6]. Our results were encouraging, with an area-under-curve (AUC) between 0.829 and 0.832 for the five trials (Table 1). The accuracy and consistency of our model shows that it can successfully identify distressed properties.

For further insights into the accuracy of our model, we also measured the accuracy of the risk scores our classifier assigned to properties. We grouped all properties from the test set into bins based on the risk score assigned to them by the classifier, and then measured the proportion of properties in each bin identified as distressed in the Neighborhood-by-Neighbor survey. Figure 3 shows the results of this analysis for one of our cross-validated models (the results for each trial of cross-validation were almost identical). The size of each circle represents the number of properties in that bin. As shown in the figure, the predicted and actual probabilities of distress in the survey are almost identical. This indicates that the risk score assigned to each property by the classifier accurately identifies the probability that the property is distressed.

4.3.2 Test on a Recent Neighborhood Survey

To further test our model on more recent data, we obtained the results from a neighborhood windshield survey conducted by the Frayser Neighborhood CDC on 12,000 properties in 2013. This survey provides Memphis' only labeled property data for years after 2011, and is therefore the only data available from recent years for large-scale evaluation of our model.

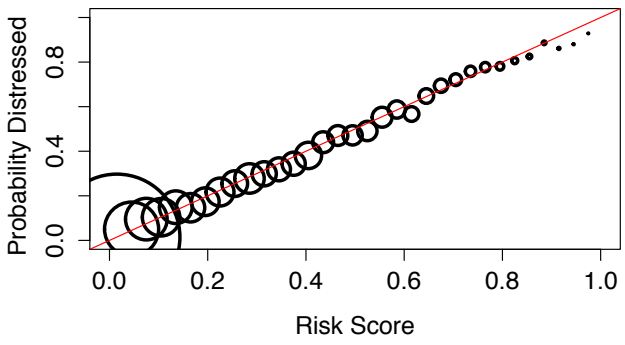

Figure 3: The risk scores assigned to properties accurately estimate the probability that the property is distressed. Circle sizes correspond to the number of properties identified with that risk score. The red line plots the one-to-one correspondence $y = x$.

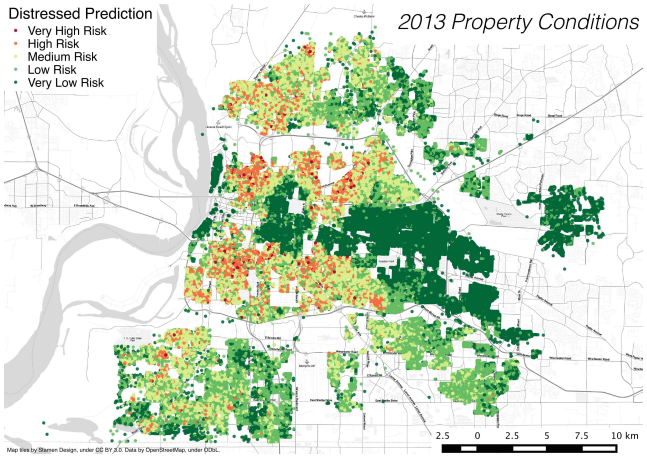

Figure 4: Our model predicts that the most severely distressed properties (marked in red) form a "C" shape around the center of Memphis.

We tested our full model on the 12,000 properties surveyed by the Frayser CDC neighborhood survey in 2013. With an AUC of 0.621, our results suggest that we are able to estimate distressed properties in 2013 with mediocre accuracy. However, we must account for a few sources of error that are likely to diminish the quality of our predictions.

Because the Neighborhood-by-Neighbor and Frayser surveys were completed by separate organizations without coordination, they label properties differently. In particular, while the Neighborhood-by-Neighbor survey identifies properties needing "cosmetic repairs only" as a distinct category (one that we considered as distressed), the Frayser survey groups properties with "possible minor cosmetic issues" and "no structural issues" in the same category (which we considered as not distressed). In other words, the Frayser survey sets a higher threshold for marking a property as distressed.

This may explain why our model significantly overestimates distress for properties it assigns a risk score between 0.4 and 0.7, but accurately predicts distress for properties with higher risk scores. Homes with "minor cosmetic issues" are likely to be classified by our model as moderate risk properties, but appear in the Frayser survey as in good condition. Our model performs better on homes to which it assigns a risk score above 0.7. These homes are likely to appear relatively consistently in both surveys. They are also, for the purposes described here, the most important properties to identify accurately. This sample of properties thus provides a better test for our model — and yields encouraging results about its accuracy.

4.4 Implementation

We used our model to estimate the condition of every residential property in Memphis in 2011, 2012, and 2013. Future years can be added to the model as new data becomes available. Figure 4 shows the estimates from our model for 2013. Note the C-shaped distribution, with distressed properties forming a strip along the north, west, and south of central Memphis. This is known colloquially in Memphis as the "C of poverty."

We have presented these estimates to policymakers and community leaders in Memphis through a web application made using the `shinyapps` package in `R` [4].[4] The site (Figure 5) provides an interactive map where users can zoom in to any neighborhood in Memphis to see each residential property's risk score. The map can display predictions for 2011, 2012, and 2013, allowing users to assess property and neighborhood conditions over time.[5]

Using this site, community groups and city officials can efficiently evaluate neighborhoods to inform investments and policy proposals. Rather than surveying the entire city to obtain estimates of which properties are distressed, they can use this website to estimate neighborhood-level priorities. Once a neighborhood has been identified for attention, the City can invest resources in preparing a much more precise and in-depth analysis of its condition.

5. EVALUATION OF REVITALIZATION STRATEGIES

Identifying distressed properties is only the first step toward revitalizing neighborhoods. The next task for cities is to determine what action to take. When dealing with distressed properties, Memphis typically either demolishes the structure, leaving a vacant lot, or boards the home, leaving an abandoned property. These actions are largely dictated by a lack of available resources. CDCs and private investors occasionally rehabilitate and sell distressed properties, inserting them back into the pool of habitable homes. While studies in other cities have linked rehabilitations with improved neighborhood outcomes [16], officials and developers in Memphis struggle to understand precisely how any individual property contributes to its neighborhood's condition.

In this section, we analyze the impact of distressed homes on neighborhoods in Memphis. Finding that rehabilitated homes are correlated with increased property values throughout their neighborhood, we then develop a tool to predict

[4]Due to restrictions on our data sharing agreement with the City of Memphis, this website is not public but the source code for all this work is available at `https://github.com/dssg/memphis-public`.

[5]We did not feel it would be useful to the City to calculate risk scores for years preceding 2011.

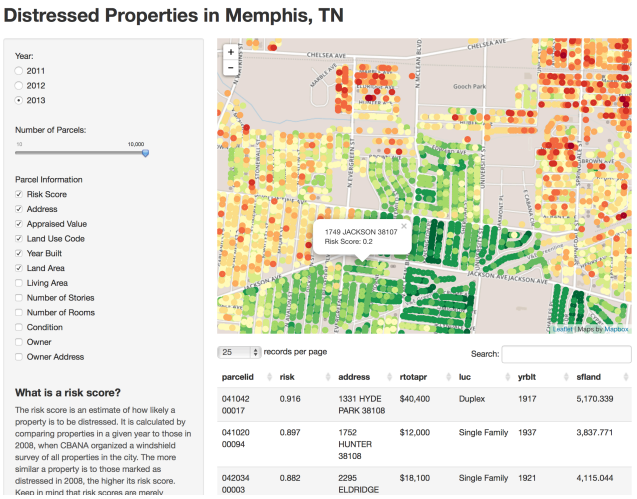

Figure 5: A screenshot of our distressed properties web application. The site contains options to control displayed data (left panel), an interactive map to look at the risk score of every property (top panel), and a data portal with information about each displayed home (bottom panel).

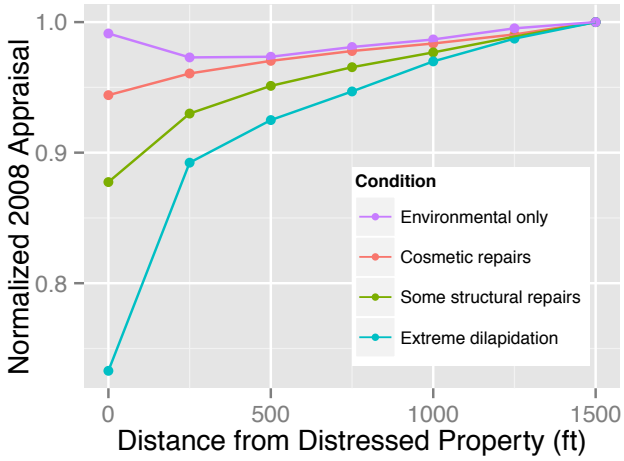

Figure 6: Homes near more severely distressed properties have lower appraised values than those further away.

which homes, if rehabilitated, will generate the greatest impacts. Together, these analyses will help Memphis design cost-effective strategies to invest in neighborhoods.

5.1 The Real Estate Market and Distress

We first studied the relationship between the property values of distressed homes and other homes in their vicinity. We selected each single-family home identified as distressed in the Neighborhood-by-Neighbor survey [5] and averaged the appraised values of the surrounding non-distressed homes as a function of distance from the distressed property. We grouped distressed properties and their neighbors based on the form of distress, as noted in the survey. We then normalized these profiles by the value of homes in the most distant bin (1250 to 1500 feet) and plotted the mean profile in Figure 6. The leftmost point is the average normalized value of the distressed properties themselves, followed by the average of properties between 0 and 250 feet away, and so on.

The results are intuitive: more severely-distressed properties and their immediate neighbors are valued below undistressed homes further away in their neighborhoods. Home values increase as the distance from a distressed property grows. Additionally, homes marked as "extremely dilapidated" show the biggest price decrease relative to their neighbors, indicating that worse forms of distress correlate with more severe reductions in home value.

It is tempting to interpret these results as an estimate of the effect of distress on neighboring home values. However, it could equally easily be interpreted as demonstrating the increased likelihood of distress in areas with unusually low property values — values whose ultimate cause may be something else entirely (crime, little accessibility to jobs, less access to credit, poor health, etc.). Moreover, distressed properties are clustered: a home near one distressed property is likely to be near others as well, which magnifies the observed price decrement. These results should therefore be taken as a characterization of the typical property values surrounding distress, rather than as a measurement of its impacts.

Another confounding factor in this analysis is our use of appraised property value as our metric of a home's worth. Assessors consider the quality of nearby properties when estimating the value of each home, and so it is difficult to untangle whether a distressed home truly impacts the value of its neighbors or simply causes assessors to provide a lower appraisal based on their formulae.

We also studied the appraised values surrounding distressed homes where some remediation was attempted. We identify rehabilitated homes as those that were labeled as distressed in the Neighborhood-by-Neighbor survey and received non-demolition building permits valued over $10,000 between 2008 and 2013. In Figure 7, we compare the neighborhoods of distressed single family homes that were demolished, rehabilitated, or received no treatment.

Unfortunately, building permits are not a perfect indicator for investment in a property. Some improvements may be done without a permit or do not require one (for example, new flowerbeds, trash pickup, or lawn maintenance). Additionally, not all permits are followed through, and some permit activity may not visibly change the house (such as interior repairs). Nonetheless, building permits are the best available proxy to infer when a property has received rehabilitation.

Using appraised values from 2013, we see that rehabilitated properties and their neighbors have a higher relative value for their neighborhood than the neighbors of untouched properties. Meanwhile, demolitions are associated with price decreases. It is not surprising that demolished properties are worth significantly less than their neighbors, since there is no structure on the lot and they were likely in the worst condition initially. The more interesting result is that the neighbors of demolished homes are appraised at lower values than the neighbors of rehabilitated homes, a result that holds at all distances considered. This suggests that rehabilitations may improve neighborhoods while demolitions drive down the value of neighboring properties.

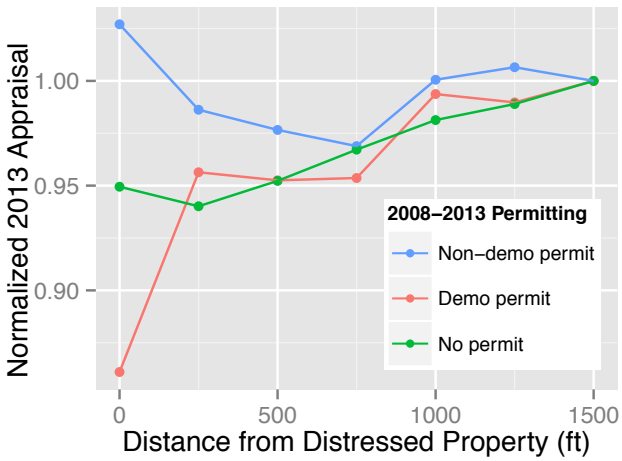

Figure 7: The appraised values of rehabilitated homes and their neighbors are higher than those of demolished homes.

These results are merely descriptive and cannot definitively evaluate the effectiveness of rehabilitation and demolition. They are designed instead to summarize how the City deals with distress and to generate hypotheses for further testing. As above, we note the complex relationships not captured in our analysis. For example, developers are more likely to invest in neighborhoods where property values are increasing, while cities are most likely to demolish properties in the worst neighborhoods. It is not yet clear what impact, if any, rehabilitating or demolishing homes has on neighborhood dynamics.

5.2 Evaluating the Impacts of Rehabilitation

While the previous analysis has shown that rehabilitations are correlated with higher property values, to guide the actions of city governments we would like to understand the causal relationships at play: do any of these interventions change the property value of homes in the neighborhood?

Understanding the indirect impacts of rehabilitations will help cities determine when interventions can be justified economically. Considering properties individually, the answer is usually no: it typically costs more to rehabilitate a property than can be recouped by its sale and increased property tax revenues. However, if the rehabilitation also increases the value of neighboring homes, then the economic calculus changes: the cumulative gains from the rehabilitated property *and* its neighbors may be greater than the cost of investment, thereby financially justifying the intervention.

Questions of causality are best answered using econometrics, controlling outside influences on the outcome of interest to isolate the effects of the variable under study. This is often done using matching, a technique in which each observation in the treatment set is matched to an observation from the control set that has similar characteristics. This method attempts to make the only difference between the treatment and control groups be the treatment itself.

In our case, we wanted to evaluate the impact of interventions on the value of neighboring homes. We looked at recently-sold single and multi-family homes that did not see building permit activity, defining those within 500 feet of rehabilitated homes as the treatment group and those more than 500 feet from rehabilitated homes as the control group. We identified rehabilitations as properties that were marked as distressed in the 2008 Neighborhood-by-Neighbor survey and had since then had been subject to building permits of $10,000 or greater.

Arguably the most common matching method is propensity score matching, in which a model (for instance, a logit) is fit to predict the probability of treatment based on the covariates. The propensity function from that model (in the case of the linear regression, the output of the linear regression before the logit transformation) is then used to match observations between treatment and control. Observations that are similar on the covariates that predict selection are matched. Thus a control group is created of samples that had a similar probability to receive treatment as those that did receive treatment.

The related technique that we used is known as proximity matching [14]. With this technique, a random forest is used to predict treatment, and the proximity matrix from that random forest is used to match observations. Thus, observations that frequently end up in the same leaves of the random forest are more likely to be matched. This shares the quality of propensity score matching wherein observations are matched primarily based on the covariates that are likely to influence selection, but also has the property that observations are matched on the covariates themselves, rather than on a linear combination. Thus one is less likely to see a result where very different observations are matched because they have unrelated traits that happen to yield similar propensity functions. We found that proximity score matching produced better covariate balance, and so we used it for the analysis that follows. We used a random forest with a minimum terminal node size of 20, 3 variables at each split, and 25,000 trees.

Recognizing that our data does not elucidate all salient attributes about each property, we attempted to match unobservable features as best we could. We controlled for location in our matching, requiring each of the control properties to be within two miles of the matched property from the treatment group. Additionally, we used neighborhood-level covariates to ensure that matched properties were in similar neighborhoods. We assume that properties are most similar to those near them, and thus that we can best control for unobserved features by selecting homes from the same region of the city.[6] Two miles is large enough to include potential matching properties, but not so large that a property could be matched with one from a vastly different neighborhood. Unfortunately there is no metric to know how well this parameter we have chosen performs, since the attributes we are controlling for are unobserved.

Performing the matching as described, using spatial restrictions and proximity matching, we found a positive effect of rehabilitations on surrounding property sales: our model estimates a 3.25% increase in property values around rehabilitated homes.[7] This number is not statistically significant

[6]This is Tobler's First Law of Geography: "Everything is related to everything else, but near things are more related than distant things" [21].

[7]The minimum p-value for differences in observed variables between treatment and control was 0.23. The minimum p-value for first-order interactions between observed variables was 0.17. These are Kolmogorov-Smirnov p-values with 1000

($p = 0.312$), however, meaning that we cannot confidently determine whether rehabilitation has an effect. Thus, the results are inconclusive.

The conclusion from this analysis should not be that rehabilitation of distressed properties does not affect neighborhood property values, but rather that we were unable to detect a statistically significant effect of rehabilitation on property values. We believe this is largely because of the small sample size of sales ($n = 200$). Houses, especially in distressed neighborhoods, don't sell very often. Additionally, the hypothesized positive effect of interventions on home values is likely small to begin with, and thus difficult to differentiate from zero effect — no one would expect to see a 50% bump in sale price due to a nearby rehabilitation.

Such insignificant results do not necessarily imply that rehabilitation has no effect on housing price. The direction of the effect is positive, as one might expect; it is simply not measurable at 95% confidence with the data we have. It is possible that collecting more data would allow us to find a stronger effect; as additional homes are renovated in Memphis neighborhoods, and more homes around them are sold, this analysis could be revisited.

One way forward would be to increase the sample size by repeating this research in a larger city, or by extending the period of sales. Another option would be to take a qualitative approach, interviewing new buyers of houses near a rehabilitation and asking questions about how the neighborhood influenced their choice, documenting the effect of the interventions directly.

Model misspecification is another possible problem. One concern is that this model assumes that rehabilitating a distressed property has the same effect on surrounding property values in every part of the city. Perhaps rehabilitation affects property values only in certain neighborhoods. Relaxing this assumption is difficult, though, as the data is sparse: there are relatively few sales around rehabilitated homes, and they do not represent every neighborhood.

5.3 Simulating Tax Appraisals

Memphis has an active set of Community Development Corporations (CDCs) that rehabilitate dilapidated properties and sell them to local residents. However, CDCs cannot rehabilitate many properties because the costs typically exceed the amount that can be recouped from the sale of the property. The Frayser neighborhood CDC, for example, typically spends about $20,000 more rehabilitating a house than it can make back in that home's sale [12].

Based on the previous analyses, however, the benefits of rehabilitating a home appear to extend beyond the sale price of that specific house. This impact can be manifested in many forms. In parts of the city with large amounts of abandoned housing, an extra resident may help stabilize a block, preventing further abandonment. Even in areas that are doing well, an extra resident may contribute to community cohesiveness. Just having a house that is physically well-maintained will reduce the perceived disorder of a neighborhood, and may increase the value of neighboring properties.

Additionally, because recent comparable sales are used to determine the assessed value of properties for tax purposes, a rehabilitation can impact the tax revenues generated by the renovated property as well as its neighbors. Unlike some of the other benefits created by rehabilitated houses, the tax impact is potentially measurable: it is computed entirely from the tax assessor's property records, a dataset made available to us.

Understanding the impacts of rehabilitation required first determining how the Shelby County Assessor of Property calculates property values. Due to external factors we were unable to correspond directly with the Assessor on this issue, but their website describes the general procedure to determine the property taxes owed by a given parcel [2]. Properties are reappraised every four years, with the most recent appraisal occurring in 2013. Prior to appraisal, an inspector visits the house to ensure that the Assessor's database reflects the current condition of the property. The Assessor also records the price and type of all property sales in the county. Once the dataset has been updated, the Assessor uses a Computer-Assisted Mass Appraisal (CAMA) system to estimate each property's value based on its condition and recent sales in the area. This is then multiplied by the assessment rate (25% for residential properties) and the result is multiplied by the tax rate (generally around 3%) to get the final tax paid each year.

We matched each of the 12,486 single family residential properties in Frayser with the five closest sales whose properties were of comparable age, size, and condition. The appraised value was computed as the average of the five prices. The correlation between our estimated appraised values for 2013 and the Assessor's appraised values was 0.74. While by no means perfect, this is high enough to suggest that it captures many of the assessor's formula's underlying mechanisms.

The ability to estimate appraised values from property and sales data allowed us to estimate what would happen to neighborhood tax appraisals following the hypothetical sale of a rehabilitated property. We simulated appraisals where the sale of a rehabilitated home is included as a comparable and one where it is not. Subtracting the second set of values from the first provides an estimate of the impact of the rehabilitation on the tax appraisal of each property. Summing across properties gives the total impact of the new appraisal in the neighborhood.

This will help investors evaluate potential rehabilitations based on their expected impact on property values. To facilitate these estimates, we created a web application that allows users to select a property in Frayser and input a hypothetical selling price (Figure 8). The website then calculates the expected impact that selling the home at that price would have on the total property tax assessed throughout Frayser.

With this tool, the Memphis government and Frayser CDC can better predict their returns for rehabilitating and selling different homes. If they use it to target investments such that they can recoup more money, their model would become more financially stable. If investors and developers did not lose money on most of their rehabilitations, they would have the funds to rehabilitate more homes, thus turning revitalization into a self-sustaining venture.

6. IMPACTS IN MEMPHIS

Our analyses have contributed to Memphis' understanding of neighborhood vitality. By providing a macro view of distress across the city, officials can now see citywide patterns and trends. We have also provided new analyses re-

bootstraps. Indicator variables used a paired sample t-test. Balance was evaluated using the `Matching` package in `R` [19].

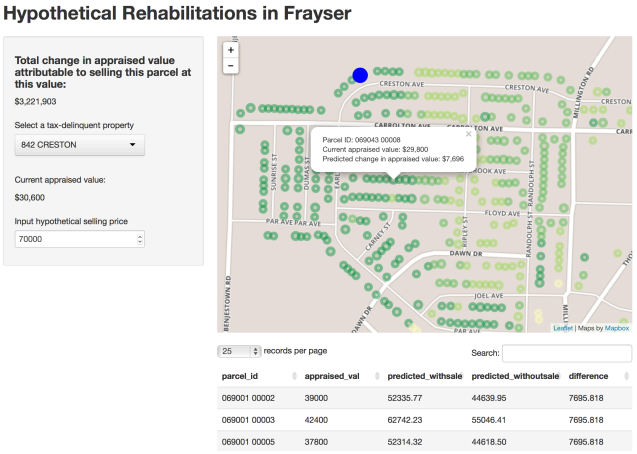

Figure 8: We built a website that predicts the impact of hypothetical rehabilitations on property values for one neighborhood in Memphis.

garding the cost of distress on neighborhoods. As such, the City is better prepared to calibrate neighborhood improvement strategies with a better sense of how it will affect a neighborhood rather than a single parcel.

As Memphis undertakes future revitalization efforts, it can do so with greater ease and to greater effect than was possible before. Its first task is to identify neighborhoods in need of aid. In previous years this would have required expensive and time-consuming neighborhood surveys just to determine which neighborhoods should be evaluated further. This can now be done online using our website described in Section 4.

Once it has decided where to invest in revitalization efforts, the City must determine what actions to take. While it has traditionally boarded or demolished distressed homes, our analyses in Section 5 challenge the suitability of these strategies. Although our results were not entirely conclusive, we found correlations between demolitions and decreased property values in the neighborhood. Rehabilitations, on the other hand, were correlated with increased property values for both the rehabilitated home and its neighbors.

To aid the rehabilitation efforts of City Hall and local CDCs, we developed an application that predicts the impact of proposed rehabilitations on neighborhood property values. They can use this tool to invest in homes that will generate the greatest increase in property values across the entire neighborhood. This will drive down the effective cost of each rehabilitation, freeing up resources to be used for additional rehabilitations or other revitalization efforts.

Our project has also started conversations locally about the need to invest money and human resources into the City's use of data. In order to aid future work by the City of Memphis, we created and handed over a data portal that contains all of the data we gathered from various sources. This affirmed the potential in Memphis for a unified data center that would allow stakeholders to be on one platform and ask questions using common data. The City is currently designing a more sophisticated data warehouse. In addition, officials around Memphis are now exploring new projects that use data to address their needs.

7. DATA SCIENCE CHALLENGES IN GOVERNMENT

Cities across the country are increasing their use of data and technology to guide policy and better interact with citizens [8, 22]. Throughout our work, we encountered many challenges that we believe are typical of efforts to apply data science techniques to aid government policy. We review several here and discuss important steps to improve future data-driven efforts in public policy.

First and foremost, we encourage cities to invest in data collection efforts: data mining algorithms produce results only as good as the data they are given. Cities and citizens should develop common goals for data-driven initiatives. This will involve first identifying problems that data can solve, then determining what data is needed, and finally developing tools to acquire and maintain data.

Our work also highlights the need for municipal bodies to share data while solving complex urban problems. One of our primary challenges was discovering what data exists, identifying who in Memphis controls it, and determining how to access it. We believe that creating a common data warehouse across City Hall will empower government employees to develop more comprehensive data-driven policies.

Another challenge we encountered was determining the proper metrics to evaluate the success of revitalization initiatives. Cities must juggle numerous, often competing, goals: balancing the budget, improving social connectivity, and ensuring the welfare of their most disadvantaged citizens, among others. Because cities are responsible for the well-being of their citizens, they cannot think like a corporation and focus only on earning back the money spent. If a revitalization project increases property values but displaces many citizens through gentrification, we do not believe that should be considered a success. Yet it was difficult to develop metrics for social vitality based on the data in city ledgers. We believe it is important for governments, data scientists, and social scientists to think creatively about novel ways to measure social well-being with existing data as well as new data that should be collected toward this end.

Finally, we encourage data scientists working with public officials to emphasize that their models provide *estimates*, and should be used as one of multiple inputs into the decision making process. There are many caveats that need to be specified (noisy data, sample bias, model assumptions, etc.), and it is important to be open about the limits of our techniques. Therefore, our results came with the attached notice: "Apply local knowledge before drawing conclusions." It is a remark equally about the limits of modeling as it is a reminder to be mindful of the context of one's work.

8. FUTURE WORK

There are multiple approaches we could take to improve our model of distressed properties. Initial attempts should focus on feature selection, in particular generating more complex features from the data. One example would be to consider the number of unique homebuyers, rather than total transaction volume, thus separating the effects of speculation from other types of investment.

Another potential approach is to take advantage of large online resources with images of city streets, such as Google Street View. This technique has recently gained attention as an effective way to study urban conditions [10, 17]. Com-

puter vision algorithms could identify features of the homes labeled as "distressed" and "not distressed" using historical images available on the site. Incorporating images into our model would add a rich layer of information about each home not captured in the City's internal databases.

Finally, it will be valuable to attempt similar applications using more targeted machine learning approaches. In the absence of up-to-date labels of property conditions, we could design a semi-supervised system where only some of the properties require training data.[8] This would ease the burden on officials or volunteers who otherwise would need to assess every property. It would also be worthwhile to attempt multi-class learning methods that can classify homes into more nuanced categories than just "distressed" and "not distressed."

9. CONCLUSIONS

In this paper, we developed tools to aid Memphis, TN's efforts to revitalize neighborhoods. Our methods were successful at identifying distressed properties based on the City's administrative data, allowing stakeholders in Memphis to easily, quickly, and inexpensively evaluate citywide trends and neighborhoods in particular need of aid. We then studied the impacts that distressed properties have on their local neighborhoods. Finding that rehabilitations appear more effective than demolitions at revitalizing neighborhoods, we built a tool to help CDCs and other local investors identify which homes will most affect their neighborhood if rehabilitated.

By using data that municipalities across the United States already collect regularly for administrative purposes, we highlighted approaches that are possible in many cities without requiring extensive upfront data collection. We hope that these analyses will show municipalities how utilizing their existing data and increasing future data-collection efforts can help them better serve citizens and use their resources to the greatest effect. Our work has, in part, spurred Memphis to invest more heavily in data collection and analysis, as well as policymaking based on data processing. This will allow both data scientists and city employees to conduct more rigorous and impactful analyses in the future.

10. ACKNOWLEDGMENTS

This work was completed during the 2014 Eric and Wendy Schmidt Data Science for Social Good Summer Fellowship at the University of Chicago. We are grateful to the entire team of mentors and fellows who contributed their guidance, suggestions, expertise, and camaraderie.

We received invaluable data and assistance from many people and organizations in Memphis. These include George Lord and Tommy Pacello from the Innovation Delivery Team, Nathan Ron-Ferguson and Tk Buchanan from the University of Memphis, the Frayser and Binghampton CDCs, Shelby County ReGIS, and the Shelby County Register of Deeds.

11. REFERENCES

[1] City-data. `http://www.city-data.com`. Accessed: 2015-02-15.

[2] Shelby county assessor of property. `http://www.assessor.shelby.tn.us`. Accessed: 2014-07-21.

[3] Urban areas in the United States: 1950 to 2010. `http://www.demographia.com/db-uza2000.htm`. Accessed: 2015-02-15.

[4] J. Allaire. *shinyapps: Interface to ShinyApps*, 2013. R package version 0.3.63.

[5] T. Buchanan, P. Betts, J. Gilman, and R. Brimhall. Neighborhood-by-neighbor: A citywide problem property audit, 2010.

[6] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[7] G. Galster, P. Tatian, and J. Accordino. Targeting investments for neighborhood revitalization. *Journal of the American Planning Association*, 72(4):457–474, 2006.

[8] S. Goldsmith and S. Crawford. *The Responsive City: Engaging Communities Through Data-smart Governance*. John Wiley & Sons, 2014.

[9] I. Goldstein and C. S. Closkey. Market value analysis: Understanding where and how to invest limited resources. *St. Louis, MO: Federal Reserve Bank of St. Louis*, 2006.

[10] J. Hwang and R. J. Sampson. Divergent pathways of gentrification racial inequality and the social order of renewal in Chicago neighborhoods. *American Sociological Review*, 79(4):726–751, 2014.

[11] A. Liaw and M. Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

[12] S. Lockwood. Personal Communication, Aug 2014.

[13] C. Marohn. Strong towns report for the City of Memphis, 2013.

[14] D. I. Miller, H. Tan, and J. Savage. The next generation of matching methods for causal inference. In Preparation.

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[16] Research for Democracy. *Blight Free Philadelphia: A Public-Private Strategy to Create and Enhance Neighborhood Value*. Eastern Pennsylvania Organizing Project, 2001.

[17] P. Salesses, K. Schechtner, and C. A. Hidalgo. The collaborative image of the city: Mapping the inequality of urban perception. *PLoS ONE*, 8(7), 2013.

[18] R. J. Sampson. *Great American City: Chicago and the Enduring Neighborhood Effect*. University of Chicago Press, 2012.

[19] J. S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: The Matching package for R. *Journal of Statistical Software*, 42(7):1–52, 2011.

[20] The National Vacant Properties Campaign. Vacant properties: The true costs to communities, 2005.

[21] W. R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, pages 234–240, 1970.

[22] A. M. Townsend. *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. WW Norton & Company, 2013.

[8] We thank an anonymous reviewer for this suggestion.