Reducing Data Poverty in NYC: Achieving Open Data for All

A Capstone Report

Prepared for:
NYC Mayor's Office of Data Analytics
August 2016

Prepared by:
NYU Center for Urban Science + Progress
MSc Applied Urban Science and Informatics Capstone Team
Ramda Yanurzha | Jiheng Huang | Maria Filippelli | Adhlere Coffy | Arno Amabile

NYU Faculty Advisors: Dani Hochfellner | Logan Werschky

Project Brief

New York City's (NYC) Open Data platform hosts over 1,500 datasets of administrative information collected by the City and covers nearly all dimensions of the city (NYC Open Data website). This data is made publically accessible online for free through the NYC Open Data Initiative, led by the Mayor's Office of Data of Analytics (MODA) (NYC Mayor's Office of Data Analytics, 2015). MODA's goal is for all New Yorkers to use and benefit from Open Data (NYC Mayor's Office of Data Analytics, 2015).

With MODA's vision of Open Data for All and the increasing amount of data available, questions arise: What does Open Data use look like in practice? Are there areas in the city that have less data than others? Or, are there certain groups that use Open Data at a disproportionately high rate? These questions address a larger concept called Open Data Poverty, a condition in which one is deprived of the benefits of Open Data due to the lack of access, use, and representation within data. The goal of this project is to quantify Open Data Poverty and investigate how to decrease Open Data Poverty in NYC. We perform an empirical analysis to measure how a lack of resources in and data about a community influences Open Data use at a community level.

Acknowledgements

The New York University (NYU) Center for Urban Science + Progress (CUSP) Open Data Poverty Capstone team would like to acknowledge and express our appreciation for the crucial role MODA staff played in the development of this project and report. Specifically we would like to thank Lindsay Mollineaux, MODA Deputy Chief Analytics Officer, and Amen Ra Mashariki, MODA Chief Analytics Officer for their guidance and support. Additional thanks goes out to Pratap Jadhav and Clare Zimmerman of Socrata for their expertise, support, and guidance while aiding us in the retrieval of use analytics for the NYC Open Data Portal.

Introduction

Open Data initiatives are being implemented at an increasing rate throughout various levels of government worldwide (McGee & Edwards, 2016). The benefits of these initiatives include increased government transparency, citizen participation in government, and interagency operability as stated in the New York City (NYC) Open

Data law, Local Law 11 of 2012. The delivery of NYC Open Data is guided by the 2015 Open Data for All strategic plan and data is provided to the general public through a web portal (NYC Mayor's Office of Data Analytics, 2015).

Increased availability of data does not automatically translate into increased benefits. It also does not address how benefits, increased or not, are distributed across members of the public. Research demonstrates that expecting the "average internet user and citizen" to access and use Open Data amounts to "wishful thinking" because not everyone has the knowledge and skills to make use of the data for specific ends (Gurstein, 2011). In addition, there are gaps in understanding how much and to what extent citizens not only use data, but also how the data represents citizens. Even with the appropriate infrastructure and skills, individuals might not be able to use Open Data effectively (Janssen, Charaalbidis & Zuiderwijk, 2012). Our project takes a unique approach to addressing these gaps within open data and, more specifically, NYC Open Data. The space where lack of skills meets lack of representation is where we find open data poverty.

We define Open Data Poverty as the situation in which one is deprived of the benefits of Open Data driven by the lack of access, use, and representation within data. With this concept in mind, we developed a series of investigatory questions:

- Why is it important to measure Open Data Poverty?
- How do we measure Open Data Poverty?
- Are there trends in Open Data Poverty across New York City?

Open Data Poverty evaluation does not currently exist, and this project will provide a unified, cohesive approach to addressing NYC Open Data Poverty quantitatively and qualitatively. We used Structural Equation Modeling (SEM) to quantify Open Data Poverty in NYC and created a web-based map tool for further qualitative exploration.

Research Design

This project tackles a topic that is largely undefined. It requires that the team first identifies what Open Data Poverty means conceptually before attempting to measure it. Current dialogues and debates in civic technology and in citizen engagement and empowerment primarily center around "The Digital Divide" and have only recently begun to touch upon assessing Open Government Data portals. Discussions identifying how disparities in data production and collection can lead to the state of being "data poor"

are few, but are the most critical to the development of ideas surrounding Open Data Poverty (Castro, 2014).

Our approach was divided into two major phases: defining Open Data Poverty and quantifying Open Data Poverty. Defining Open Data Poverty includes the study of different factors that which represent this concept. We refer to them as individual Actors – the acting forces that influence Open Data use. An integral step in the identification and selection of the individual Actors was to define the geographic level of analysis. The guiding ideas used were the availability of data, the ability to act on insights gained from the analysis, and the accuracy of the data that would be collected at the geographic level capturing how the user is influenced by that specific Actor. With these considerations in mind, the team selected Community Districts as a workable geographic level of interest and as a single body of individuals with common goals and concerns with respect to the public good.

We hypothesize that six major Actors contribute to Open Data Poverty: demographics; awareness; empowered organizations; community concerns; infrastructure; and representation.

To quantify our assumptions we exploit a mathematical approach called Structural Equation Modeling (SEM). This enables us to reveal the system of relationships between Open Data Poverty and its contributing Actors, as shown in Figure 1. Open Data Poverty is measured as an inverse to Open Data “use” and is evaluated quantitatively to examine the correlation between Open Data Poverty and Actors.

An SEM enables testing of complex multivariate relationships between different variables of interest that cannot be measured directly (Weston and Gore Jr., 2006). An SEM proceeds in two phases: (i) a Measurement Model, in which a metric is created for each Actor based on an array of different variables; and (ii) a Structural Model, in which the relationships between Actors are estimated numerically.

A Measurement Model consists of aggregating different variables into a composite score for each Actor (Wittkowski et al., 2004). Thus for our model, we collect data for each Actor (Demographics, Representation, Infrastructure, Community Concerns, and Empowered Organizations). Data variables were identified through client interviews and literature review, and capture dimensions of Actors as that directly influence Open Data use.

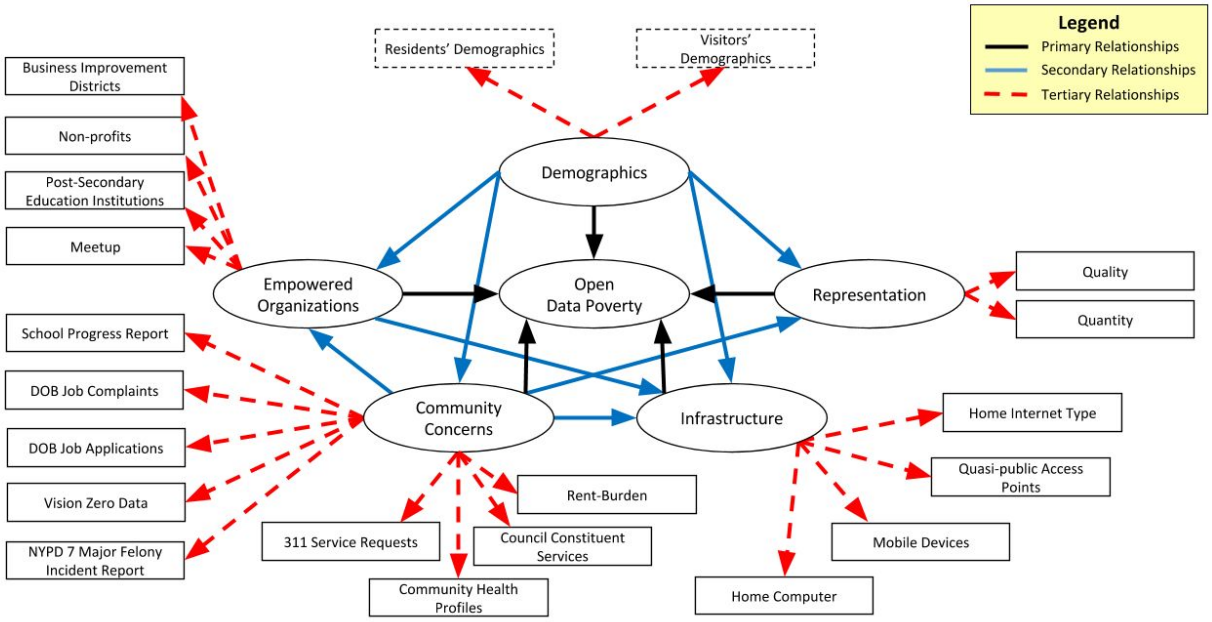


Figure 1 - Open Data Poverty Systems Diagram

In addition, we conducted secondary research to understand how each Actor interacted with other Actors to approximate the use of Open Data in NYC and measure Data Poverty.

System Actors

The five system Actors – Community Concerns, Demographics, Empowered Organizations, Infrastructure, and Representation – comprise, along with Open Data Poverty, the structural model. Each of these Actors, as well as Open Data Poverty, are latent, and thus cannot be measured directly. This section details the manifest (i.e. measurable) variables that are used to compute the composite score of each Actor, which by proxy produce an approximation of Open Data Poverty.

Community Concerns

Two main sources – Community District websites and a custom Community Board member survey – inform the approach to quantify community concerns at the Community District level. A Community District website review revealed that only two

community districts had a reference to NYC Open Data, though most had multiple links to city data. The most common city data references were 311, Community Health Profiles, and school information. The Community Board survey provided insight to community data needs by asking Board members about datasets they download, main community concerns, and data needs for existing problems. The survey revealed the top datasets requested and used to be 311, Department of Buildings (DOB), and NYPD information. Other frequently referenced datasets included housing, and traffic. Full survey results can be found in Appendix A.

Some Community District websites also display data from NYC Open Data. We combined the survey results with the information displayed on Community District websites for a fuller picture of community concerns. Based on the frequency in the survey and on Community District websites, top concerns include: Department of Buildings information such as job applications and job complaints; the School Progress Report for educational information; safety concerns including Vision Zero data and the NYPD 7 Major Felony Incident reports; Community Health Profiles; and reported concerns such as those listed in the 311 and Council Constituent datasets. All datasets are available on the NYC Open Data Portal. Community Concern summary statistics are shown in Figure 2. The minimum display the community district with the least concerns, the maximum shows the community district with the highest number of concerns. They reveal that many variables have a right skew as evidenced by the average of most variables being larger than the median. The DOB data has the largest skew, the Community Districts with the most permit applications and complaints are more than five standard deviations away from the median.

COMMUNITY CONCERNS									
	Recorded						Reported		
	7 Major Crimes	DOB Permits	Rent Burden	Health Profile	School Report	Traffic Accidents	Council Contacts	311 Service Requests	DOB Complaints
min	615	228	37	2	40	1,117	47	4,620	1,094
max	4,641	8,574	64	19	69	6,234	3,409	22,467	15,310
median	1,786	768	52	10	49	2,745	428	11,529	3,576
average	1,811	1,146	52	11	51	2,921	600	11,057	3,869
std dev	709	1,252	7	5	6	1,207	549	3,040	2,036

Figure 2: Community Concern Summary Statistics

Demographics

Comparing communities includes a comparison of their demographic make-up. Indicators such as job type and age provide insight on social profile of a community.

This project also considered data on population, income, language and education of community districts.

People employed in certain industries use data on daily basis and have more interactions with Open Data. Better data allows financial service providers to reduce costs and deliver better quality services to consumers (Castro, 2014). Insurers can pose new questions and better understand many different types of risks, with better access to open government statistics and third-party data from a wide variety of sources (Clarke & Libarikian, 2014). Job type data in this project is collected from U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) data, including Residence Area Characteristics (RAC) and Workplace Area Characteristics (WAC). People working in three industries − 'Information', 'Finance and Insurance', and 'Professional, Scientific and Technical Services' − are included in this data collection process.

Age is also relevant. People age eighteen to twenty-nine and age thirty to forty-nine are among the age groups that use open government information the most (Horrigan & Rainie, 2015).

Among other datasets, population data provide a baseline when analyzing demographic components in a community districts and comparing similarities and differences between them in New York City. Income is also an important demographic indicator as it has an impact on the 'digital divide'. People earning higher income or living in affluent communities have more access to infrastructure and services to use open data (Castro, 2014). And fluency in English has a large positive effect on immigrants' earnings in the United States (Bleakley & Chin, 2004). Finally, people with college degrees are more engaged in using Open Data to monitor government performance (Horrigan & Rainie, 2015).

The statistical summary in Figure 3 shows that community district vary significantly on number of workspace employments in the selecting industries (information, finance, insurance, professional, scientific and technical services). The Bronx community 1 has the smallest number of people with bachelor degree or higher. Median population with bachelor's or above degrees is 31,471. Brooklyn community district 13 has the least population age eighteen to forty-nine at 39,299 and Manhattan community district 12 has the largest at 117,171. Median household income data are collected in this project. The most affluent area is Manhattan community district 2 West Village, Greenwich Village and SoHo area. The median annual household income in this area is 120,190, almost 6 times the lowest amount in New York City, which is the Bronx community district's 20,872 dollars per year. Queens community district 14 has the largest

population that speak English as native language or speak Spanish as native language but also speak English very well. Finally, on population, the Bronx community district 8 has the smallest population in New York City, at 106,737. Queens community district 7 has the largest at 256,742.

DEMOGRAPHICS SUMMARY STATISTICS							
	Population	Income	Age	Job Type		Education	Language
				Resident	Visitor		
min	106737	20872	39299	2160	294	8798	36821
max	256742	120190	117171	39306	495266	143746	173488
median	150100	50316	76677	10250	2299	31471	90307

Figure 3: Demographics Summary Statistics

Empowered Organizations

Making government documents accessible through New York State's Freedom of Information Law of 1974 did not in itself increase government transparency. This required the production and mobilization of resources by organizations (Harisson, Pardo & Cook, 2012). Similarly, Open Data use requires more than just publishing data on a platform. A network of organizations "empowered" by Open Data is necessary to connect people with technical skills or domain knowledge on given challenges (Janssen, Charalabidis & Zuiderwijk, 2012), to distribute funding through grants, to provide prizes that may attract funding, and to attract public attention (whether by diffusion of information created through Open Data or by adoption of an application relying on Open Data).

There are several organizations composing this ecosystem. Some offsprings of the Open Data movement are created with the explicit purpose of using or incentivizing the use of Open Data. Others are established community organizations formerly focused on public outreach. For Juan Camilio Osorio, Director of Research at the NYC Environmental Justice Alliance (NYC-EJA), "[empowered] organizations are part of community-based planning and can do research provided with the right technical and financial resources" (Sanborn-Hum, 2016).

The presence and activity of empowered organizations in Community Districts thus play a crucial part in using and promoting Open Data. Research on the types of organizations known to use or sponsor use of Open Data leads us to consider Business

Improvement Districts (BID), Meetup groups, post-secondary education institutions, and nonprofits.

BIDs are public-private partnerships designed to maintain, promote, and develop a commercial district. As such, they have sought to spur the creation of services based on Open Data for their member businesses or customers: they have provided data for the BigApps competitions in New York and sponsored a mapping of parking rules in New York City. Because we cannot report all cases of New York City BIDs using or sponsoring Open Data, we consider that the presence of a BID signals the existence of a potential sponsor or user, and a certain willingness of businesses to mobilize resources, potentially for Open Data use. For each community district, we count the number of businesses belonging to a BID with Open Data datasets from the New York City Open Data.

Open Data Meetup groups are relatively informal groups gathering volunteers to work on challenges with Open Data. In New York, the largest is BetaNYC. But it not alone − 155 Meetups groups are related to Open Data in New York City. As those different groups may have varying levels of activity, we scraped the number of Meetup events by community district.

Secondary education institutions, with their concentration of funding and skilled people (both in technical and domain expertise), are considered as one of the key elements of Open Data ecosystems. Various research projects at the NYU Center for Urban Science Progress and the Columbia Center for Spatial Research use Open Data and hold resources that can (and are) mobilized locally. Given that institutions vary in their size, we compute the total secondary education population at the community district, from a federal dataset.

Finally, as underscored by Juan Camilio Osorio, non-profits have the potential to be significant actors in the ecosystem. For instance, non-profits often contribute to BigApps − the biggest competition based on Open Data in NYC − as a subject matter expert or community resource. Without data on their mission statements or their activities, we rely solely on the number of non-profit organizations by community district though a New York State dataset.

Summary statistics show that 43 community districts contain a Business Improvement District, 16 do not. In Community Districts containing a BID, the number of businesses in a BID varies from 4 to 2,847 (Community District 5 in Manhattan, Midtown), with an average of 430 businesses in a BID per community district. There have been Meetup

events related to Open Data in 29 community district, 30 having hosted no Meetup events from a group devoted to Open Data. On average, a community district that hosted at least one Meetup event hosted 53 of them, but 50% of the community districts that hosted at least one Meetup event hosted less than 6 of them. There are campuses of higher education in 43 community districts, 16 having no such campuses. The smallest higher education "population" (students + researchers) in a community district is 108 (Community District 9 in Queens), the maximum is 81,820 (Community District 2 in Manhattan, around Greenwich Village & SoHo). On average, community district that host a higher education campus have a higher education population of 16,233. There are nonprofits filed in 54 community districts, 5 community districts having no filed nonprofits. In community districts where there is at least one nonprofit, the number of nonprofits filed varies from 1 to 9,513 (Community District 5 in Manhattan, Midtown). Community districts, when they host a nonprofit, host on average 627 of them.

Infrastructure

Infrastructure is critical to understanding the distinction between the potential for Open Data use and what Open Data use actually looks like, making government portals not simply a mechanism for transparency but a tool for improving governance. Without physical access to Open Data portals, via hardlines or WiFi connections, the promise of Open Data is severely limited and hinders uptake by those most likely to benefit from it. We identify infrastructure by four major components: Home Internet Type, Computer Use, Mobile Devices, and Quasi-Public Access Points.

Computer Use and Mobile Device demonstrate the separate importance of each mode of accessing the Internet. The rise in ubiquity of mobile phones, tablets, and other devices grant citizens extensive access to the Internet but limits utility as these devices are geared for social media, not data analysis. Computer Use in households illustratives demand for the ability to interact with platforms with additional functionality that isn't available on mobile devices (Rosston, Savage & Waldman, 2010).

Quasi-Public Access points, including libraries, WiFi hotspots, and LinkNYC Kiosk are government run or supported resources that provide individuals an opportunity to access the internet free of charge. Some services provide adequate bandwidth but are limited in geographical reach (Becker et al., 2011).

Figure 4a and 4b outline summary statistics of the measurements for infrastructure. Of the 59 Community Districts, only 5 have less than 75% of the households with a computer. These 5 Community Districts are BX 3, 4, & 6, MN11, and BK13. MN08 has

the largest number of households at 114,838 with 94% having a computer. The average and median for the all Community Districts is ~83% demonstrating a fairly consistent use of computers in households. The number of households that have a mobile subscription is considerably lower than expected. MN08 has the highest percentage of households with a mobile subscription at 55%. 41 Community Districts have less than 1/3rd of the households reporting have a mobile subscription. Of the 3 groups for Home Internet Type, the percentage of households that report no Internet access varies considerably across all Community Districts. MN 1 & 2 have less than 8% of the households reporting no Internet access while BK16 registers the highest with nearly 40% of the households not having Internet access. The average for all Community Districts is 20%. Access points vary greatly across all Community Districts. Despite each community district having at least two library access points, the majority of WiFi hotspots are in Manhattan. All Community Districts in Manhattan have multiple WiFi access points with MN10 having 113, the most of any Community District in the city. This is compared to Staten Island that has only 2 between all 3 districts or Brooklyn that has only 3 districts with WiFi points. BK2 has 103 while BK 3 & 6 have 1 each, the rest have zero. LinkNYC kiosks are only in Manhattan.

Infrastructure Summary Statistics					
	Households with Computer	Households without Computer	Households with Mobile	Households without Mobile	Number of Libraries
min	30869	3859	5927	24508	2
max	108746	18612	63973	65785	12
median	42576	8892	15239	39287	4

Figure 4a: Infrastructure Summary Statistics

Infrastructure Summary Statistics					
	Number of WiFi Hotspots	Number of LinkNYC Kiosk	Households with High Internet Connection	Households with Low Internet Connection	Households with No Internet Connection
min	0	0	24666	1339	4225
max	113	50	98992	8144	22408
median	1	0	36643	3679	11465

Figure 4b: Infrastructure Summary Statistics

Representation

The concept of Representation is based on the idea that individuals or groups ‘demand’ Open Data and are more likely to use it when data is relevant to the individual’s or group’s interest. The usefulness and importance of any data is through the perceived value of those directly using the data. Data representation is when data captures some aspect of the data user, their community, or neighborhood. Without a sense of being ’counted’ − represented − in public data, users are less likely to actively seek out datasets. In this research, representation is viewed through a geospatial lens and defined as the occurrence of an individual record in a dataset that can be georeferenced via latitude and longitude coordinates or street address. The use of a spatial component, whether as a primary feature or supporting information, has been a hallmark of hyperlocal news reporting in the context of civic engagement. Engaged citizens tend to value spatially relevant information as it represents them and their communities (Metzgar, Kurpius & Rowley, 2011).

While New York City Open Data portal encompasses a large variety of data, some datasets such as 311 Service Requests are examples of georeferenced data that play a significant role in providing spatially-relevant information on a community. New platforms such as neighborhoods.nyc and datausa.io, in turn, create new ways for citizens to better understand their neighborhood characteristics and help drive civic participation in their community.

To measure this spatial representation, we developed a simple geolocator that maps entries containing geographical information to its corresponding Community District in New York City, with a minimum of complete address and the borough where it is located. While many high quality commercial geocoder exists, our sample size is in the magnitude of millions of entries that makes such endeavour to be economically prohibitive and time consuming. The simplicity and strictness of our geolocator also serves as a litmus test that indicates ideal georeferencing - complete parsability of a complete address means that it is a valid address in the official map of New York and can be reliably mapped by more complex solution. This property is especially important as reliable address parsing is still a computationally demanding problem due to the large variety on how it is written and area-specific edge cases. To determine the presence of this geographic variable, we used probabilistic parser combined with manual inspection to identify and label variable fields on each dataset in NYC Open Data portal. We purposefully excluded 311 Service Request Dataset and very large ones such as TLC Taxi Trip dataset as they are already used in other actors,

computationally expensive to calculate, or does not fit with the concept of data poverty in general.

From 1,025 dataset in the portal, we identified 124 data file in CSV format with variables containing at least full address and borough identifier with a total of 4,968,626 rows. Of this, we successfully parsed 3,480,864 (56.7%) valid addresses with their corresponding community district. The parsability ratio of each dataset ranges from 0 to 0.93 and generally does not have a direct relationship with its size. As low parsability ratio can be caused by more fundamental issues on how geographic information is encoded in the data, we used 0.5 as a threshold to exclude data having too many geolocation failure, which results in a final observation of 73 datasets with average parsability ratio of 0.77. From this per-dataset result, we computed two metrics: representation occurrence and quality. While occurrent is simply a sum of geolocated rows for each community district, quality is a weighted per community district sum of each dataset's parsability ratio multiplied by the sum of geolocated rows with an average of 0.85.

Open Data Use

To observe Open Data Poverty the team determined that Open Data use signals the absence of Open Data Poverty. To infer Open Data Poverty, the team measured the number, frequency and location of NYC Open Data downloads. The team collaborated with Socrata in order to obtain Socrata site analytics as well as Google Analytics information. The data provided a number of variables, ranging from unique users by dataset to the number of times a specific dataset was downloaded or viewed. Despite the richness of the metrics retrieved, variables were only available at the city level and not at the Community District level. City-wide data is not useable in SEM analysis; however portal site use informed inferences about Open Data use at the Community District. We used the presence (binary) and proportion of the number of rows where a given community district is included in each dataset multiplied by its corresponding download number as a variable representing open data use.

Methods

Measurement Model - Composite Scores

The first step of an SEM is to estimate the Measurement Model, which consists of creating a metric for each Actor based on the variables defining them:

1. Organizations: A high score captures a Community District with a relatively high number of businesses belonging to a BID, many nonprofits, a large population working or studying in post-secondary institutions, and many Meetup events related to Open Data.

2. Representation: A high score captures a Community District for which there is a high quality and quality of parseable geo-spatial information in Open Data. That is, there is enough data on location to identify from which district the data comes.

3. Demographics: The score captures both the demographics of the resident population and the working visitor population (people working in the Community District without necessarily residing in it). A high score for the resident demographics captures a Community District where there are many 18-54 years old residents, with a relatively high income, who speak English, have a college education, are rather male and white, and hold technical jobs. A high score for the working visitor demographics capture a Community District in which jobs are rather technical, held by young and educated workers.

4. Community Concerns: A high score captures a Community District in which there are many concerns, both recorded by city services (many traffic accidents, high crime level, poor health profile, many building permits, many rent-burdened households), or reported by people (to constituents services, to 311, to DOB for building complaints).

5. Infrastructure: A high score captures a Community District where there are many homes that have computers connected to the Internet via broadband, a mobile subscription, and access to publicly provided WiFi.

The distribution of those scores is plotted and shown in Figure 6.

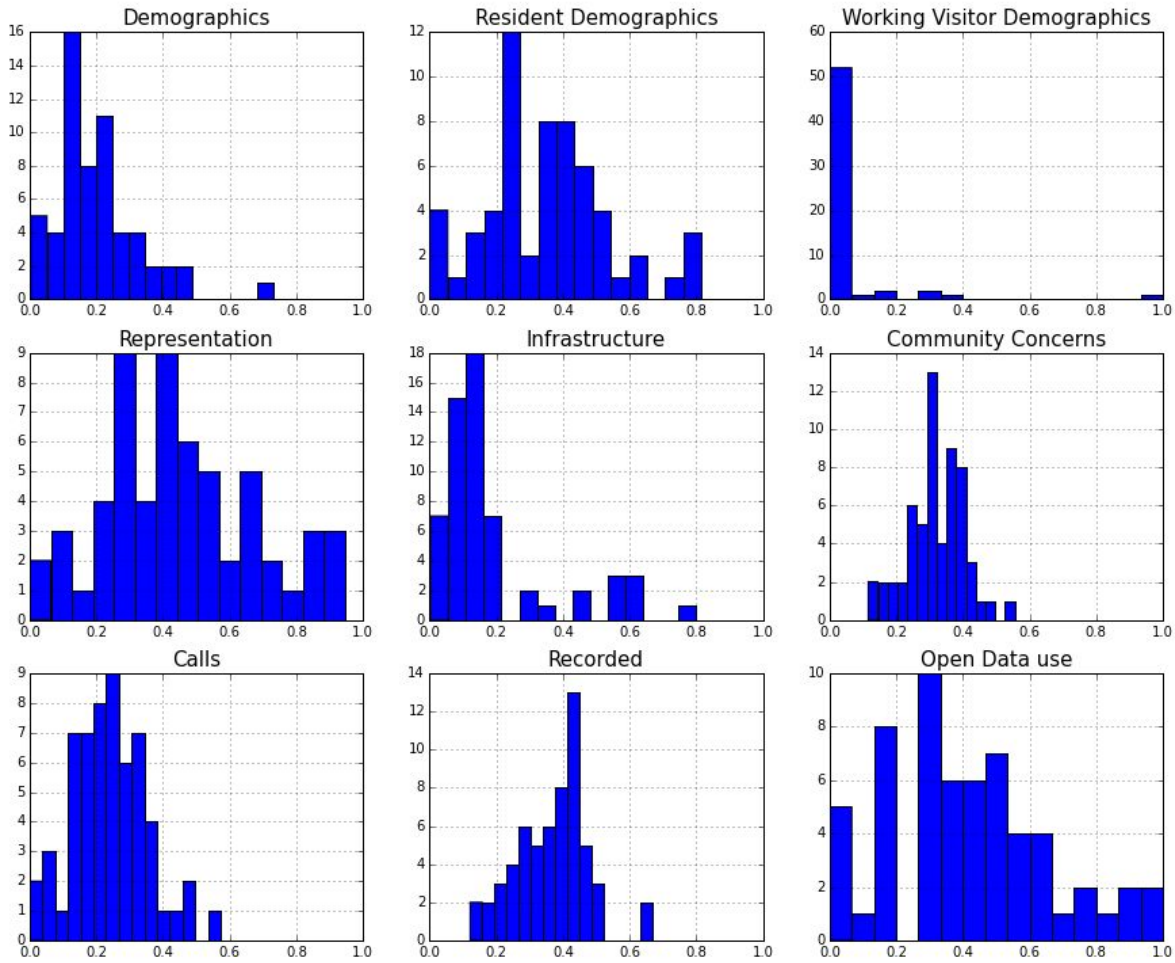


Figure 6: Distribution Plots of System Actors

Structural Model - Equations

With each Actor assigned a composite score, we then estimate the Structural Model. Due to the complexity of SEMs, the relationships between Actors using a Structural Model should be developed theoretically, and the number of specifications should be limited to a handful of variations (Weston & Gore Jr., 2006).

We test four different specifications of the SEM, all defined by a set of six equations with a different Actor for each equation as the “dependent” variable. As this report focuses on the results of the analysis, we focus here on the Representation Actor and Open Data Poverty (measured with Open Data use). The complete table of results for all specifications is presented in Appendix B.

Specification (i)

(i) The "initial" specification, corresponding to the System Diagram, where the Open Data Poverty Actor is measured by its inverse, Open Data use (ODU):

$$ODU = f(ORG, INFR, REP, CON, CON^2, DEM) \;\&$$
$$REP = f(DEM, CON, CON^2)$$

Specification (ii)

(ii) We test the additional effect that Infrastructure may have on Representation:

$$ODU = f(ORG, INFR, REP, CON, CON^2, DEM) \;\&$$
$$REP = f(DEM, CON, CON^2, INFR)$$

Specification (iii)

(iii) We split the Demographics Actor (DEM) into its part capturing the profile of residents (RES) and its part capturing the profile of working visitors (people who work in the CD but do not necessarily reside in it - VIS):

$$ODU = f(ORG, INFR, REP, CON, CON^2, RES, VIS) \;\&$$
$$REP = f(RES, VIS, CON, CON^2, INFR)$$

Specification (iv)

(iv) We split the Demographics Actor between residents and working visitors, and split the Concerns Actor (CON) into its part capturing concerns recorded by the city (RECO), and concerns reported by constituents (mainly calling - CALL):

$$ODU = f(ORG, INFR, REP, CALL, CALL^2, RECO, RECO^2, RES, VIS) \;\&$$
$$REP = f(RES, VIS, CALL, CALL^2, RECO, RECO^2, INFR)$$

Results

Throughout the modeling portion of this analysis, the only metric that is consistently found to have a statistically significant relationship with Open Data use is the Representation composite score (which is always significant at the 1% level). Given the coefficient range of Representation (0.36-0.42), an increase of one standard deviation

for the Representation composite score is associated with an increase of the Open Data use composite score between 0.085 and 0.111, on average, ceteris paribus. The increase in Open Data use resulting from this increase in Representation represents between 37% and 48%, depending on the specification in question, of the standard deviation for the Open Data use composite score, thus a sizable effect.

This finding yields two interpretations. First, the inference can be made that the Actors identified for the system are either insufficient in one of three ways:

1. The variables suffer from lack of robustness as metrics for measuring Open Data use, e.g. the number of nonprofits filed in a Community District may be a very poor proxy for the activity of nonprofits in that Community District;

2. The variables are too heterogenous, e.g. the combination of home internet connection and quasi-public access points yields an effect that is different at the Community District level than what it could be on a household level (Vaupel and Yashin, 1985).

3. Or, lastly, the variables are entirely irrelevant to measuring Open Data use.

Second, it seems however that the quantity and parsability of the data for a given Community District on the Open Data platform is relevant for Open Data use. This finding is very encouraging and demonstrative of the novelty of the work conducted in this analysis. It substantiates one of the most original hypotheses of this analysis – that is, as Representation within the data increases for a given Community District the Community District's Open Data use will also increase by some magnitude. There is, however, the possibility that given the way in which the composite score for both Open Data use and Representation is computed, their estimated relationship might be driven largely by the similarity in mathematical modeling rather than an underlying statistical phenomena.

Given the importance of Representation, evaluating the relationship between this Actor and the other Actors in the model can provide insight. Results from specifications (i) and (ii) suggest that:

1. A Community District with a higher score for Demographics is expected to have a lower score for Representation (statistically significant at the 1% level);

2. A Community District with a higher score for Community Concerns is expected to have a higher score for Representation (significant at the 5% level);

3. The score for Infrastructure is not significantly related to the score for Representation indicating no statistical connection between a Community District's ability to access the Internet and its representation in Open Data use.

A higher composite score for Demographics captures both a younger, more educated resident population with higher income and a more educated, distinctly large number of tech-oriented employees that don't live in the Community Districts. On the other hand, a higher score for Community Concerns captures both a Community District with more concerns as recorded by City services and reported by constituents through 311 calls. The heterogeneity of those Actors leads us to redefine their composition to dig deeper and uncover the dynamics at play.

Results for specifications (iii) and (iv) suggest that the relationship between Representation and Demographics in a Community District is primarily driven by the Demographics of Working Visitors (coefficient statistically significant at the 1% level, at the 10% level only for Residents). An increase in one standard deviation of the composite score for Working Visitors is associated with a decrease of the Representation composite score by 0.162 (69% of the standard deviation of the score for Representation), on average, ceteris paribus.

Conversely, the relationship between Community Concerns and Representation is driven by the concerns recorded by City services (coefficient significant at the 10% level, not significant for concerns reported through calls). An increase in one standard deviation of the score of Concerns recorded by City services, when the score of around the mean, is associated with an increase, on average, of the Representation composite score of 0.168 (71% of the standard deviation of the score for Representation), ceteris paribus. The inclusion of the squared term for Community Concerns suggest that there are "diminishing returns" of the amount of Community Concerns on Representation.

Visitors' Demographics and Recorded Concerns thus seem to have a significant relationship with Representation. It is likely that the more problems a Community District has, the more incidents are logged by City services which will result in an increase in the amount of data available for the Community District on the NYC Open Data platform.

Implications

There are two ways to better understand how to make the results applicable to Community Districts, through an interactive map tool created for this project and case studies. A web-based map tool has been created specifically for this project and further exploration of the results. The tool, entitled “NYC Open Data Poverty”, can be found at bit.ly/datapovertymap. A screenshot of the tool’s landing page is shown in Figure 7.

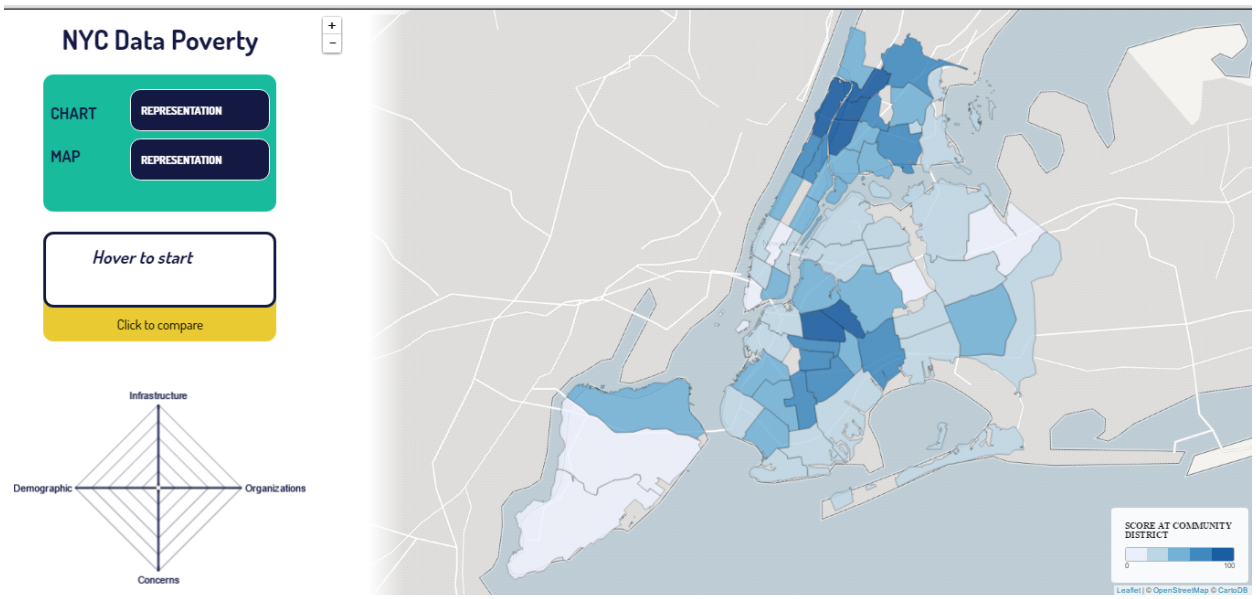


Figure 7: NYC Open Data Poverty Map Landing Page

Two case studies were selected for further analysis, Queens Community District 12 (QN12) and Bronx Community Districts 1 and 4 (BX01 and BX02).

QN12 has a moderate level of data representation and the adjacent districts (Queens Community Districts 8, 9, 10, and 13) are more than 40% different in the representation metric. The neighborhoods of Jamaica, Hollis, St. Albans, Springfield Gardens, Baisley Park, Rochdale Village, and South Jamaica constitute QN 12, which is located on the eastern central side of Queens. Diving into the details of why QN12 is different, we find that that it has a higher Community Concerns score from the adjacent Community Districts, which coincides with the model results. Specifically, this Community District has higher crime, more indicators of poor health, a greater rent burden, and lower school metrics than the adjacent districts.

However, QN12 also proves some of our initial assumptions may not have been correct. The higher proportion of residents aged 18-49, native English speakers, and males than the adjacent districts would initially lead us to believe that there are less Community Concerns than the adjacent districts. Since this is not the case, demographics may not have the negative effect on representation that we originally hypothesized.

The Community District of BX01 is comprised of the Mott Haven, Port Morris, Melrose. Adjacent to the north of BX01 is BX04. The neighborhoods of Highbridge, Concourse, Mount Eden, and Concourse Village comprise BX04. BX01 and BX04 are very similar to each other in terms of demographics, infrastructure, and many of the variables that constitute the Community Concerns metric. This would lead us to believe that the representation metric should be similar. However, this is not the case and the Representation metric has a difference of 70%. Diving deeper into this metric we found that BX04 has approximately 70% more representation in three datasets, Housing Maintenance Code Violations, Rodent Inspection, and Housing Maintenance Code Complaints. Digging even further into this discrepancy, we found that BX01 has a larger supply of public housing than BX04. Because public housing records are not posted on the NYC Open Data portal, this could be one reason for the difference.

The case studies provided are examples of ways to interpret the results and use the map tool. We provided one way to strictly interpret the NYC Open Data system using QN12 and its adjacent Community Districts. Another way to use the tool is as a starting point for further investigation as shown with BX01 and BX04.

Conclusion and Further Use

This project brought to light a new definition of Open Data Poverty, the condition in which one is deprived of the benefits of Open Data driven by the lack of access, use, and representation within data. We treated Open Data Poverty as the intersection between the Digital Divide and the Data Divide measuring unique components of each to socio-technical phenomenon.

Increases in Open Data use resulting from increases in Representation represented between 37% and 48%, depending on the specification, of the standard deviation for the Open Data use composite score, thus a sizable effect.

As Representation proved to be the most important Actor in terms of Open Data use, we evaluated the statistically significant relationships Representation had with other Actors. The relationship between Representation and Demographics in a Community District is primarily driven by the Demographics of Working Visitors. Conversely, the relationship between Community Concerns and Representation is driven by the concerns recorded by City services.

Our analysis shows above all things that an empirical analysis of the effects of socio-technical divide phenomena can be conducted using Open Data resources and yield statistically significant results. Although these results are heavily dictated by the choice of geographics scale, availability of data, and heterogeneous influences of the data, our project provides a sound proof of concept that future empirical analyses can build upon to expand the catalogue of work in this arena.

Further use of our project should rely on the interactive map tool to further investigate the relationships and target community outreach to increase representation in Open Data. All materials of this project can be accessed at the CUSP Data Facility https://datahub.cusp.nyu.edu/.

References

Attard, J., F. Orlandi, S. Scerri, and S. Auer (2015). A systematic review of open government data initiatives. *Government Information Quarterly,* 32(4), 399-418.

Becker, S., M. D. Crandall, K. E. Fisher, R. Blakewood, B. Kinney, and C. Russell-Sauvé (2011). Opportunity for all: How library policies and practices impact public Internet access. *IMCS-2011-RES-01*. Institute of Museum and Library Services. Washington, D.C.

Bleakley, H., and A. Chin (2004). Language skills and earnings: Evidence from childhood immigrants. *Review of Economics and Statistics*, 86(2), 481-496.

Castro, D. (2014). The Rise of Data Poverty in America. *Center for Data Innovation,* September 10.

Clarke, R., and A. Libarikian (2014). Unleashing the Value of Advanced Analytics in Insurance. *McKinsey & Company*, August.

Chaudhuri, A. (2005). An analysis of the determinants of internet access - ScienceDirect. *sciencedirect.com/science/article/pii/S0308596105000674*

Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday,* 16(2).

Harrison, T. M., T. A. Pardo, and M. Cook (2012). Creating open government ecosystems: A research and development agenda. *Future Internet,* 4(4), 900-928.

Horrigan, J., and L. Rainie (2015). Americans' Views on Open Government Data. *Pew Research Center.*

Janssen, M., Y. Charalabidis, and A. Zuiderwijk (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management,* 29(4), 258-268.

McGee, R., and D. Edwards (2016). Introduction: Opening Governance–Change, Continuity and Conceptual Ambiguity. *IDS Bulletin*, 47(1).

Metzgar, E. T., D. D. Kurpius, and K. M. Rowley (2011). Defining Hyperlocal Media: Proposing a Framework for Discussion. *New Media & Society,* 13(5), 772–87.

NYC Mayor's Office of Data Analytics (2015) "Open Data for All". Retrieved from http://www1.nyc.gov/assets/home/downloads/pdf/reports/2015/NYC-Open-Data-Plan-2015.pdf

NYC Open Data. January 2012. https://nycopendata.socrata.com/

Pew Research Center (2015). Americans' Views on Open Government Data. Retrieved from pewinternet.org/2015/04/21/open-government-data

Rosston, G. L., S. J. Savage, and D. M. Waldman (2010). Household demand for broadband Internet in 2010. *The BE Journal of Economic Analysis & Policy*, 10(1).

Sanborn-Hum, K. (2016). New York City's Evolving Approach to Open Data. *Gotham Gazette, April 11.* Retrieved from gothamgazette.com/index.php/city/6272-new-york-city-s-evolving-approach-to-open-data.

Savage, S. J. (2009). Ability, location and household demand for Internet bandwidth. *sciencedirect.com/science/article/pii/S0167718708000702*

Socrata (2011). 2010 Open Government Data Benchmark Study. Retrieved from socrata.com/benchmark-study

Vaupel, J. W., and A. I. Yashin (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician,* 39(3), 176-185.

Weston, R. and P. A. Gore Jr. (2006). A Brief Introduction to Structural Equation Modeling. *The Counseling Psychologist*, 34(5), 719-751.

Wittkowski, K. M., E. Lee, R. Nussbaum, F. N. Chamian, and J. G. Krueger (2004). Combining several ordinal measures in clinical studies. *Statistics in Medicine,* 23(10), 1579-1592.

Appendix A:

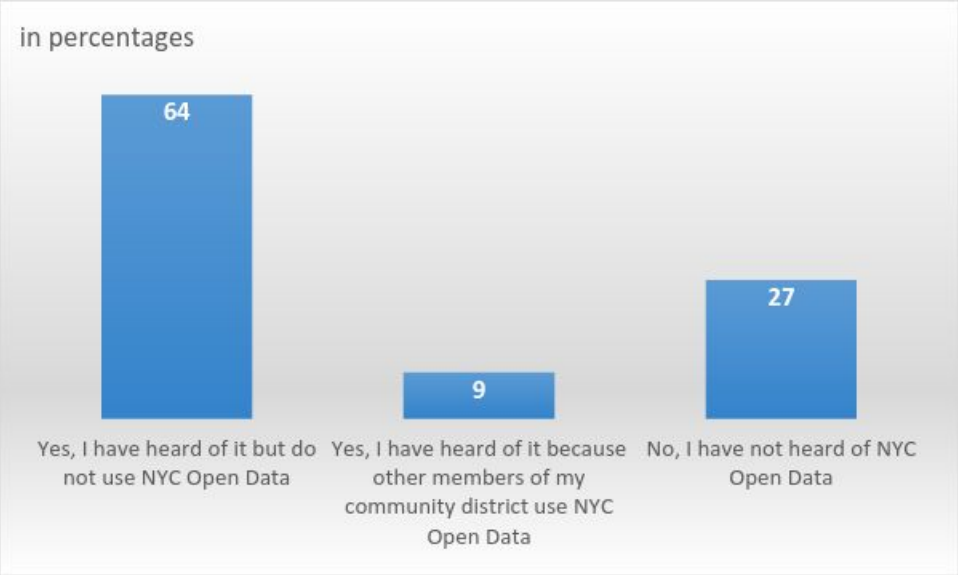
COMMUNITY BOARD MEMBER SURVEY SUMMARY

Q: Please let us know which borough you serve.

Borough	Number of Responses	Number of Unique Community Districts
Brooklyn	3	3
Bronx	9	1
Manhattan	7	3
Queens	13	4
Staten Island	0	0

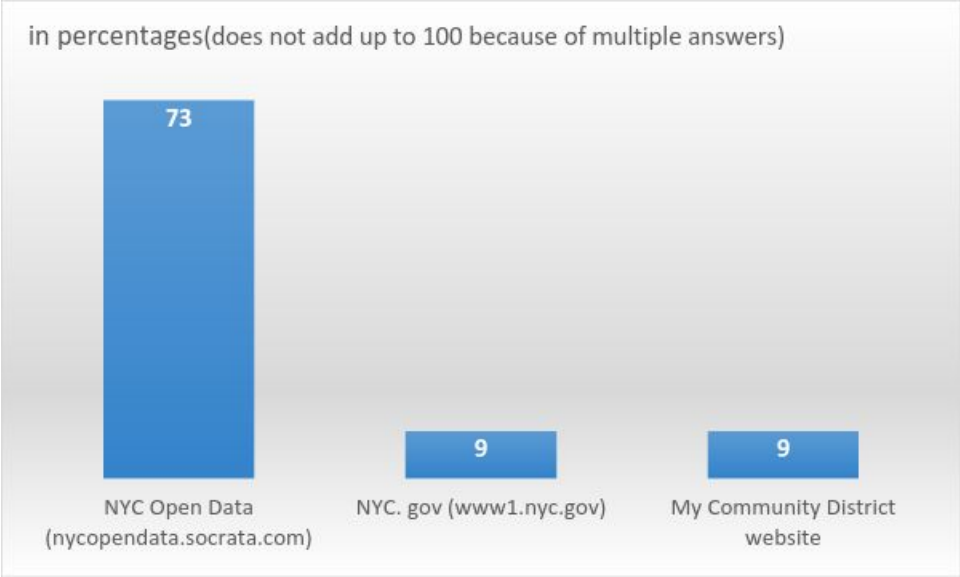
The following analysis is done one the community district level, which means that community district for which we had multiple survey responses are combined into one community district cell. This means we have data on 11 Community Districts.

Q: Have you heard of NYC Open Data?



Yes = 8, No = 3

Q: For those that said they have heard of NYC Open Data, the site is accessed by:



When asked to list the top 3 datasets most typically used or downloaded, the responses were (most popular first):

311/complaint/service request data
Demographics/census data on housing and other socio-economic information
NYPD datasets
Department of Buildings
Other – Parks, information for seniors, DOHMH, schools, work permits, political boundaries, MMR, SAPO, ECB

When asked about specific topics sought on NYC Open Data, about 63% of the Community Districts seek data on specific topics. When asked list the top 3 topics you typically seek data on, the responses were fairly similar (most popular first):

311/complaints

Demographics/census
Housing
Buildings
Schools
Parking
Other – senior centers, Quality of Life, Emergency Room information, pedestrian ramps

About half of the Community Districts participating in the survey said that members of their district had approached them with information created from NYC Open Data

When asked "To the best of your knowledge, do you know where NYC Open Data is being used in your community?", we only received information on two Community Districts:

Community groups (both tech-friendly and not)
Individuals
Tree count
311
Non-profits
City agencies
Schools

The Community Districts that had not heard of open data were asked how likely they were to use it in the future on a scale of 1 (not likely) to 10 (definitely likely): in the aggregate all Community Districts were neutral. When asked if their community had a general need for city data, they all answered with yes, but complained that information is not easily accessible.

All respondents were asked to describe their main community concerns, the top five categories are:

transportation (includes traffic and parking)
schools/education
housing (building and affordability)
development/overdevelopment
crime (gangs, violence, etc)

When asked about other data needs, the top responses were:

311/complaints
health
economics
development (unspecified)
more information on COMSTAT
street safety

Survey is available to view at:

https://docs.google.com/forms/d/e/1FAIpQLScqJLxY-yqOr5c6py-N-E2_dhsSagUBH3oiSQp1njvIoFuppA/viewform

Appendix B:

EQUATIONS

(note: The Roman numerals next to each equation correspond to the specifications in the DETAILED MODEL RESULTS table)

Representation

$REP = f(DEM, CON, CON^2)$ (i)

$REP = f(DEM, CON, CON^2, INFR)$ (i)

$REP = f(RES, VIS, CON, CON^2)$ (iii)

$REP = f(RES, VIS, CALL, RECO)$ (iv)

Infrastructure

$INFR = f(DEM, ORG, CON, CON^2)$ (i) & (ii)

$INFR = f(RES, VIS, ORG, CON, CON^2)$ (iii)

$INFR = f(RES, VIS, ORG, CALL, RECO)$ (iv)

Community Concerns (CON) - Split between Concerns recorded by City services (RECO) & Concerns reported by constituents (CALLS)

$CON = f(DEM)$ (i) & (ii)

$CON = f(RES, VIS)$ (iii)

$RECO = f(RES, VIS)$ & $CALL = f(RES, VIS)$ (iv)

Organizations (ORG) - Split between Residents (RES) and Visitors (VIS)

$ORG = f(DEM, CON, CON^2)$ (i) & (ii)

$ORG = f(RES, VIS, CON, CON^2)$ (iii)

$ORG = f(RES, VIS, CALL, RECO)$ (iv)

Open Data Use (ODU)

$ODU = f(DEM, REP, INFR, CON, CON^2, ORG)$ (i) & (ii)

$ODU = f(RES, VIS, REP, INFR, CON, CON^2)$ (iii)

$ODU = f(RES, VIS, REP, INFR, CALL, RECO)$ (iv)

DETAILED MODEL RESULTS

	Representation				Infrastructure		
	(i)	(ii)	(iii)	(iv)	(i) & (ii)	(iii)	(iv)
Demographics	-0.74*** (0.22)	-0.99*** (0.37)	-	-	0.99*** (0.16)	-	-
Residents	-	-	-0.17 (0.14)	-0.38 (0.23)	-	0.52*** (0.08)	0.58*** (0.10)
Visitors	-	-	-0.65*** (0.17)	-1.06*** (0.22)	-	0.17 (0.20)	0.09 (0.21)
Infrastructure	-	0.19 (0.23)	-	0.49** (0.23)	-	-	-
Organizations	-	-	-		0.34*** (0.12)	0.61*** (0.19)	0.53*** (0.19)
Concerns	2.87** (1.45)	2.69** (1.46)	2.32* (1.43)	-	1.26* (0.78)	1.13* (0.77)	-
Concerns^2	-1.83 (2.33)	-1.36 (2.38)	-1.14 (2.28)	-	-2.79** (1.25)	-2.55** (1.23)	-
Reported concerns	-	-	-	-0.19 (0.77)	-	-	-0.12 (0.14)
Recorded Concerns	-	-	-	1.39*** (0.22)	-	-	-0.22 (0.12)*

	Concerns		Recorded Concerns	Reported Concerns
	(i) & (ii)	(iii)	(iv)	(iv)
Demographics	0.27*** (0.08)	-	-	-
Residents	-	0.19*** (0.05)	0.11* (0.08)	0.34*** (0.07)
Visitors	-	0.04 (0.07)	0.14* (0.09)	-0.13* (0.09)

	Organizations			Open Data Poverty/Open Data use		
	(i) & (ii)	(iii)	(iv)	(i) & (ii)	(iii)	(iv)
Representation	-	-	-	-	-	0.37** (0.17)
Demographics	0.91*** (0.12)	-	-	-0.45 (0.43)	-	-
Residents	-	0.12** (0.05)	0.13* (0.07)	-	-0.28 (0.22)	-0.23 (0.29)
Visitors	-	0.94*** (0.06)	0.92*** (0.07)	-	0.24 (0.45)	0.15 (0.51)
Infrastructure	-	-	-	0.16 (0.27)	0.21 (0.27)	0.23 (0.31)
Organizations	-	-	-	0.14 (0.27)	-0.25 (0.43)	-0.16 (0.46)
Concerns	-1.03 (0.82)	-0.11 (0.52)	-	-1.57 (1.65)	-1.59 (1.63)	-
Concerns^2	1.03 (1.32)	-0.16 (0.83)	-	3.92* (2.63)	3.81* (2.61)	-
Reported concerns	-	-	-0.05 (0.10)	-	-	0.07 (0.32)
Recorded Concerns	-	-	-0.13 (0.08)	-	-	0.67* (0.37)