Comparative Study: Reducing Cost to Manage Accessibility with Existing Data

Claire Chu
Southern Methodist University, chuc@mail.smu.edu

Bill Kerneckel
Southern Methodist University, bkerneckel@mail.smu.edu

Eric C. Larson
Southern Methodist University, eclarson@lyle.smu.edu

Nathan Mowat
Southern Methodist University, nmowat@mail.smu.edu

Christopher Woodard
Southern Methodist University, christopherw@mail.smu.edu

Follow this and additional works at: https://scholar.smu.edu/datasciencereview

Part of the Architectural Engineering Commons, Business Administration, Management, and Operations Commons, Business Law, Public Responsibility, and Ethics Commons, Civil and Environmental Engineering Commons, Computer Engineering Commons, Cultural Resource Management and Policy Analysis Commons, Management Information Systems Commons, Management Sciences and Quantitative Methods Commons, Public Affairs, Public Policy and Public Administration Commons, Statistics and Probability Commons, Strategic Management Policy Commons, Technology and Innovation Commons, Tourism and Travel Commons, Urban, Community and Regional Planning Commons, and the Urban Studies and Planning Commons

Comparative Study:
Reducing Cost to Manage Accessibility with Existing Data

Claire Chu[1], Bill Kerneckel[1], Dr. Eric Larson[1], Nathan Mowat[1], Chris Woodard[1]

[1] Data Science @ Southern Methodist University, Dallas
Texas, United States of America
{ChuC@, BKerneckel@, ECLarson@lyle., NMowat@, ChristopherW@}smu.edu

Abstract. "Project Sidewalk" is an existing research effort that focuses on mapping accessibility issues for handicapped persons to efficiently plan wheelchair and mobile scooter friendly routes around Washington D.C. As supporters of this project, we utilized the data "Project Sidewalk" collected and used it to confirm predictions about where problem sidewalks exist based on real estate and crime data. We present a study that identifies correlations found between accessibility data and crime and housing statistics in the Washington D.C. metropolitan area. We identify the key reasons for increased accessibility and the issues with the current infrastructure management system. After a thorough explanation of the datasets used, we also delve into some of the important variables and their meanings. We investigate how crime and housing data can be used as a means to predict possible accessibility issues. We compared our sidewalk rating predictions generated by the crime and housing data to the ratings generated by "Project Sidewalk". Using random forest modeling of local area real estate pricing and crime, we predicted the sidewalk accessibility issues better than random chance. We present the findings and discuss possible explanations for notable correlations. After thoroughly exploring our results, we investigate future enhancements of the research. The results will help city planners and policy makers more efficiently allocate infrastructure budget for sidewalk accessibility, not only in the Washington D.C. area, but in other cities as well.

1   Introduction

Washington D.C., our nations capital, is the epicenter of American tradition. With all Federal Government Agencies interspersed between national monuments and foreign embassies, it is imperative that Washington D.C. maintains a forward-looking stance with regards to both infrastructure and technology. Fellow computer science students at the University of Maryland utilized this drive to develop "Project Sidewalk", a unique web-based application that allows users to rank sidewalk accessibility around Washington D.C. Ultimately, "Project Sidewalk" aims to generate a complete and up-to-the-minute mapping of all accessibility issues so that wheelchair and mobile scooter users are better able to plan their route around the city. After reviewing their data and findings, we would like to further extend the reach of this project by

investigating possible correlations between problematic sidewalk infrastructure, real estate, and crime data.

Using crime data from June 2016 to June 2017 from crimemap.dc.org and real estate data from June 2013 to May 2017 from zillow.com, we completed random forest analysis to predict accessibility issues. We used the "Project Sidewalk" data to verify our findings, with the results showing an accuracy score that was better than just merely selecting problem sidewalks by chance. We used a geometric grid scale to granulize the data block by block and compared the results to the neighborhood grouped housing data and found that regionality played a part in the results. In general, the fewer the grid squares, the better the accessibility prediction. For nearly all our predictions, we had an accuracy of at least 60% or higher. For future research, we plan to investigate sidewalk traffic patterns to further optimize our results. With the problem sidewalk predictions and a ranking of most frequented sidewalks, like those surrounding public transportation or near tourist attractions, city planners can efficiently allocate a budget to fix accessibility problems.

By using this data, cities are able to use real estate and crime data as a means to get accessibility data without spending exorbitant amounts of money. We want problematic public infrastructure to be fixed proactively, instead of retroactively, as it is now. Creating a more efficient system to identify and eventually prioritize sidewalk problems is the future of public infrastructure and using the methods described here is a low cost, accurate way to accomplish that. We have documented the relevant findings from "Project Sidewalk", why they are relevant, a brief overview of the data we utilized, the results from our testing and finally, our analysis regarding the Washington D.C. accessibility data, crime statistics, and housing prices. We plan to use data from "Project Sidewalk", "crimemap.dc.gov", and "zillow.com" to run analysis and determine potential relationships between sidewalk accessibility, crime statistics, and real estate valuations in order to efficiently predict accessibility data.

2 "Project Sidewalk", Accessibility, and Washington D.C.

"Project Sidewalk" was created by team of students from the Human-Computer Interaction Lab at the University of Maryland lead by Dr. Jon Froehlich. Designed to establish a live mapping database for wheelchair and mobile scooter users, we wanted to expand the usage of this data. Currently, no efficient methods exist for the city to evaluate quality of sidewalks and communicate problems to the appropriate maintenance group. Walking through all sidewalks manually would be time consuming and lack real-time cohesiveness. The existing system consists of citizens reporting problems, a surveyor assessing the problem, and then, when and if the budget is available, the sidewalk is fixed. Based on this system, we can see that fixing these sidewalks is a retroactive activity. We advocate a proactive stance that allows for city planners and policy makers to better allocate time, money, and resources to the areas we predict will be problematic based on the data we analyze here. Issues like the lack of wheelchair ramps, uneven pavement, and impediments like trashcans, telephone poles and trees can affect the accessibility of a city for pedestrians, disabled citizens and potential travelers. Improving these can increase the public usage of

sidewalks, reduce injuries, and attract additional visitors to the Washington D.C. area. [16]

2.1 The Importance of "Project Sidewalk"

The knowledge of accessibility barriers for people with ambulatory disabilities in the built environment around them is hugely beneficial. As technology advances and the collection of data becomes more tractable, people with mobility impairments can evaluate built environment accessibility easier and plan more efficient travels.

"Project Sidewalk" documents accessibility in Washington D.C. and aims to leverage the data into a real-time mapping software to allow for wheelchair and mobile scooter users to determine safe travel routes. In "Project Sidewalk", end users review live images from Google Street View and evaluate street intersections and affiliated sidewalks for curb ramps, no curb ramps, obstructions, occlusions, and surface problems in 1000 feet sections at a time. As of June 2017, this data set has grown to over 70, 000 labels and 500 miles covered, or about 50% of Washington D.C. sidewalk ratings completed [16]. "Because labeling sidewalk accessibility problems is a subjective and potentially ambiguous task," the design of the app was developed through extensive research with mobility impaired persons and thus, the targeted end users are these mobility impaired persons [7]. As the project progressed and a standard rating system emerged, the developers trained "Amazon Mechanical Turks" to help assist with rating the rest of the sidewalks in Washington D.C. Now, the labeling app is open and available to the public and with the completion of a short training session, anyone can help flag sidewalk accessibility problems. This helped ensure that the quality of the ratings throughout the dataset were consistent enough for our analysis. The beauty of "Project Sidewalk" is that it leverages free data that already exists. Because it uses images from Google Street View as a basis for the labeling, the users are able to submit sidewalk labels at anytime, from anywhere.

Additionally, without "Project Sidewalk", accessibility data collection was expensive, because it would require large teams to cover the area in entirety, and inaccurate, since the time it would take to record all the sidewalk statuses exceeds the time it would take to fix them. Because of this, there was never a desire in the public sector to develop this type of data collection. Since the data from "Project Sidewalk" was low-cost, it opened up doors to look into other uses for the accessibility data.

Fig. 1. Example of Problem Sidewalks and how they are labeled in "Project Sidewalk" [16]

2.2 Accessibility Benefits

While the ultimate goal of collecting accessibility data is to ensure that everyone has equal access to the same public infrastructure, there are many other benefits to having a highly accessible city. Among them, the three that stood out in our research were the increased health and air quality, increased pedestrian safety, and increased livability. We found these to be important because they benefit residents, visitors, and the agencies that are footing the bill.

One of the benefits of city with a high accessibility score is that it generally results in adults that have better health, perhaps because they can walk to more places instead of drive. In an article titled *Children's physical activity and parents' perception of the neighborhood environment: neighborhood impact on kids study*, the authors reference this trend, mentioning, "that higher residential density and increased street connectivity designs are more conducive for adult physical activity" [17]. As the walkability for the city or neighborhood increases, adults are more likely to have higher levels of physical activity. Accessibility has become an important issue for city planners and public health officials alike. In the article, *The Walkability Premium in Commercial Real Estate Investments,* the authors posit that, "in order to reduce preventable cancers linked to obesity and inactivity, governments should require increased walking facilities, developers should construct more projects that promote walking, and employers should occupy buildings that facilitate physical activity. Similar goals were endorsed by former U.S. Secretary of Health and Human Services Donna Shalala in her address to the Urban Land Institute in 2006" [13]. Improvements in sidewalk infrastructure would help promote walking to and from the workplace. Additionally, an increase in pedestrian traffic reduces the number of motor vehicles on the road. This can increase air quality and promote better environmental standards for communities. The United States Environmental Protection Agency explains that the side effects of motor vehicle related pollution include increased risk of "asthma attacks, emphysema, and chronic bronchitis" [18]. Residents would greatly benefit from the healthier lifestyle and improved air quality created by increased sidewalk accessibility.

Additionally, another proponent for the collection of sidewalk accessibility data is the increased pedestrian safety. In fact, in a review of urban performance measures of Washington, DC, the authors note that, "The issue of pedestrian congestion is exacerbated and becomes more critical in the District as every year DC welcomes approximately 20 million visitors" [3]. Sidewalk inaccessibility for pedestrians, handicapped, and visitors feed into complicated pedestrian congestion patterns. Simply fixing sidewalks with obstructions or building sidewalks on busy streets that do not already have them could ease foot traffic and reduce the number of pedestrian fatalities, an important task when 20 million visitors, as well as almost 700,000 residents, are constantly coming and going. With increased sidewalk accessibility, visitors will have an enhanced travel experience and will be more likely to visit again.

Residents take advantage the resulting increased livability, as reflected in increased housing prices for areas with high accessibility. While the paper by LaCour-Little references accessibility as representing all infrastructure, including proximity to freeways and shopping areas, the author notes that "the infrastructure and services enhance the quality of life of the residents, and hence increase the demand for

housing" [9]. This is evidence that accessibility is positively correlated to housing prices. Additionally, an article by Boyle, mentions "buyers prefer pedestrian access to commercial uses and a 15.5% premium for houses in neighborhoods with new [walkability] features." [1]. Even if other methods of transportation are available, like bus, or train, homebuyers were willing to pay 15.5% more to be able to walk. Governments can increase their tax revenue and community livability by attracting more residents with walkability features.

Fixing sidewalks is a relatively low-cost way to increase accessibility that is proven to have benefits for residents, visitors, and community leaders. The overall health for a community with high accessibility is generally better than inaccessible neighborhoods. Visitors are more likely to return to cities where the pedestrian accessibility and safety is high. Governments will attract more residents based on the increased level of livability with better pedestrian access. These are important benefits that help outweigh the costs of collecting sidewalk accessibility data.

2.3   The Importance of Washington D.C.

According to the 2012 U.S. census, Washington D.C. has the second highest percentage of walking commute trips of U.S. cities at 11.9% [19]. In 2014, Washington D.C. enacted the "moveDC" plan, a multi-modal transportation improvement effort to upgrade the transportation network of Washington D.C. as a whole. The components of this plan were to develop an improved system to "prioritize sidewalk maintenance and repair" and "provide a sidewalk on at least one side of every street" [19]. Because Washington D.C. has already decided to make sidewalk accessibility a priority, we want to use our analysis to help them reach these goals. Currently, the city waits for citizens to call in and report problem sidewalks. Once the city gets a report, they send a representative out to review the issues, then the sidewalk is fixed when and if the budget allows. But, the Washington D.C. Department of Transported reported in their 2014 fourth quarter report that there was still about $27 million worth of backlog on unresolved sidewalk issues. This means that there was $27 million worth of problems reported and reviewed, but never fixed [6]. Our research provides a better means to budget and plan for sidewalk issues and prioritize their repair pattern. This allows for repairs to be planned proactively instead of the current reactive system.

3   About the Data

In an effort to make our research easily reproducible for Washington D.C. and other cities who wish to enhance their pedestrian infrastructure, we used datasets that were publicly available and regularly updated, with the exception of the "Project Sidewalk" data we used for validating our findings. A community's crime and real estate pricing are often recorded in great detail and readily available so these methods can easily be applied to other areas. In total we used four different datasets, a regional look at the

"Project Sidewalk" labels, Washington D.C. real estate pricing, a granular look at the "Project Sidewalk" labels, and Washington D.C. crime statistics.

3.1 "Project Sidewalk" Data by Neighborhood

Our first dataset, briefly shown in Table 1, was a dataset of all the "Project Sidewalk" labels aggregated by neighborhood. We started out with this dataset because it seemed like the simplest and easiest to work with. Preliminary analysis showed that the dataset consisted of about 60 neighborhoods in the Washington D.C. area and their corresponding "Project Sidewalk" flagged issues. It is important to note that some of the neighborhoods were not completely assessed. This was indicated by the "coverage" variable. From this dataset, we were able to determine the total number of problems by neighborhood for each of the four listed problem categories. Important things to note about this dataset were that there were some anomalies like the "Cleveland Park" neighborhood, which reported abnormally high numbers for all categories. We suspect this is because the starting point for the training data is in this neighborhood. Since everyone has to partake in the "Project Sidewalk" orientation before they can label any sidewalks, there is a disproportionate amount of labels generated for this area. Indeed, looking at the "coverage" we can see that only about 44% of the neighborhood has been covered so all the problem sidewalks come from less than half of the neighborhood. It is also important to note that this dataset only presented four possible sidewalk problems; "Curb Ramp", "No Curb Ramp", "Obstacle", and "Surface Problem".

Table 1. Sample of Regional "Project Sidewalk" Data

region_name	id	CurbRamp	NoCurbRamp	Obstacle	Surface Problem
Adams Morgan	198	10.4117647	0.55882	2.088235	3.3265306
American University Park	195	4.91379310	0.84482	0.206896	2.7179487
Barnaby Woods	218	2.82926829	0.90243	0.853658	2.075
Cleveland Park	261	274.964912	45.7719	2.982456	0.9545454

3.2 Washington D.C. Real Estate Pricing

This dataset, the first entries shown in Table 2, was comprised of average Washington D.C. neighborhood price per square foot monthly from June 2013 to May 2017 as documented by Zillow.com (Table 2). This dataset was important to show the pricing trends within neighborhood. While we were not specifically interested in the actually pricing, we were very interested in how the pricing fluctuated from month to month, year to year.

Table 2. Sample of the Zillow.com Data by Neighborhood

region_name	id	X2013.06	X2013.07	X2013.08	X2013.09	X2013.10	X2013.11
Adams Morgan	529	533	537	543	550	555	529
American University Park	508	512	516	520	524	527	508
Barnaby Woods	452	455	458	461	466	469	452
Cleveland Park	497	501	505	511	517	519	497

3.3 Granular "Project Sidewalk" Data

In contrast to the neighborhood view of the "Project Sidewalk" data, this dataset contained every sidewalk label and the exact latitude and longitude coordinates of the problem. As we can see from Table 3, this provided a very granular view of the data and also allowed us to spot-check the issues ourselves in GoogleMaps. This dataset consisted of almost 46,000 entries.

Table 3. Sample of Granular "Project Sidewalk" Data

TYPE	LATITUDE	LONGITUDE	LABEL	PANORAMA_ID
Point	38.83062363	-76.99840546	Other	FK6VRoD5oynoE7SFlS2bTQ
Point	38.86484146	-76.98968506	Occlusion	3D-IX4jzVTDUpdpd1sKwuw
Point	38.91714096	-77.03665924	CurbRamp	se2ms7gFJ-RjPsxCZlq9bA

3.4 Washington D.C. Crime Statistics

This data was collected from crimemap.dc.org for June 2016 to June 2017. The entries here contain data about when the crime was reported, where the offense was, what type of offense was reported, and if applicable, resolve date. This dataset, previewed in Table 4, was very extensive with nearly 38,000 entries. There were many other variables but we will focus on the ones mentioned above. It's important to note that the location of crimes is sometimes blinded due to pending litigation so addresses listed are generalized to blocks. We have converted these block coordinates to latitude and longitude to better compare them to the granular "Project Sidewalk" data. Possible offenses fall within two categories, violent crimes and property crimes. Violent crimes include homicide, sex abuse, robbery excluding gun, robbery with gun, assault with a dangerous weapon excluding gun, and assault with gun. Property crimes include burglary, theft, theft from automobile, stolen automobile, and arson. The "shift" variable refers to the responding officers assignment at the time of response. Industry standard shows the typical day shift runs from 7AM to 3PM, typical evening shift runs from 3PM to 11PM, and typical midnight shift runs from 11PM to 7AM.

Table 4. Sample of Granular "Project Sidewalk" Data

REPORT_DATE	SHIFT	OFFENSE	METHOD	ADDRESS
4/23/17 23:10	MIDNIGHT	THEFT/OTHER	OTHERS	1600 14TH STREET NW, WASHINGTON, DC
4/22/17 21:09	EVENING	THEFT/OTHER	OTHERS	1600 14TH STREET NW, WASHINGTON, DC
4/11/17 16:26	EVENING	THEFT/OTHER	OTHERS	1600 14TH STREET NW, WASHINGTON, DC

4 Results and Analysis

4.1 Regionality

Looking through the "Project Sidewalk" data by neighborhood, we were able to determine that there would need to be some analysis done at varying scales. We started with the neighborhoods because this gave us the largest regions (Figure 2). Combining the housing prices and the sidewalk labels, we ran a variety of test to establish the best statistical analysis based on accuracy. The results have been documented below and ordered by accuracy in Table 5. For each sidewalk label, we ran a K-nearest neighbors, a random forest test and train, a logistic regression, and a support vector model. Based on the accuracy results, we selected only the top 10 variables and re-ran the test to optimize the accuracy. "Curb Ramp" and "Surface Problem" both resulted in the highest accuracy scores for random forests, with "No Curb Ramp" showing random forest coming in a close second. Based on this initial analysis, we continued with only random forest modeling.
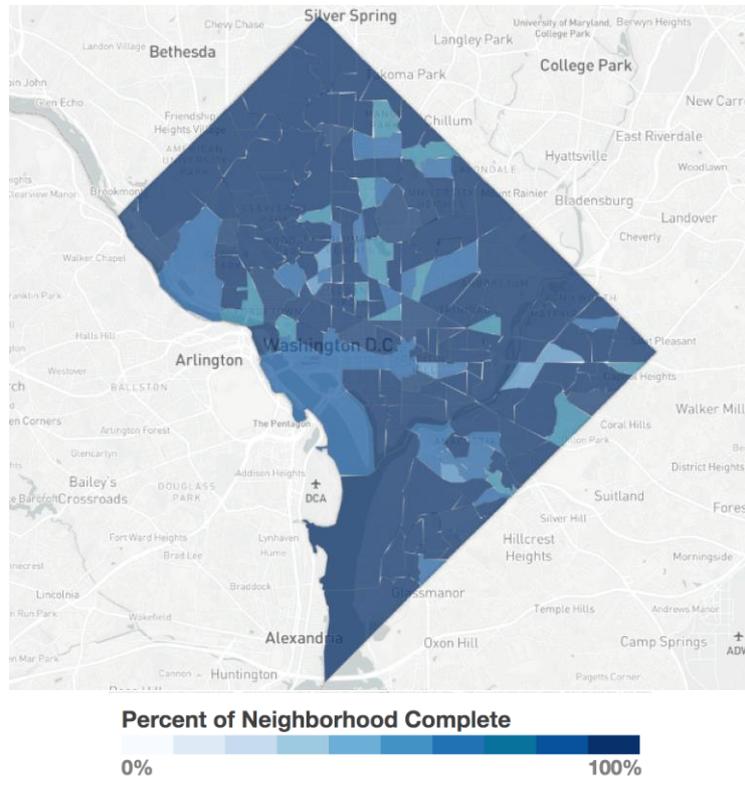
Fig. 2. Washington D.C. with "Percentage of Neighborhood Complete" Color Key as Reported by Project Sidewalk [16]

Table 5. Accuracy Results for Sidewalk Labels

CURB RAMP		
	Confusion Matricies	Accuracy
SVM	[[5 4]	0.4375
	[5 2]]	
KNN	[[23 8]	0.573770492
	[18 12]]	
Log. Reg	[[3 3]	0.625
	[3 7]]	
R.Forest Test	[[19 12]	0.639344262
	[10 20]]	
R.Forest Train	[[19 12]	**0.672131148**
	[8 22]]	

NO CURB RAMP		
	Confusion Matricies	Accuracy
R.Forest Test	[[19 11]	0.590163934
	[14 17]]	
R.Forest Train	[[21 9]	0.606557377
	[15 16]]	
KNN	[[26 4]	0.606557377
	[20 11]]	
Log. Reg	[[3 3]	0.625
	[3 7]]	
R.Forest (10)	[[19 11]	0.672131148
	[9 22]]	
SVM	[[3 3]	**0.75**
	[1 9]]	

OBSTACLES				SURFACE PROBLEM		
	Confusion Matricies	Accuracy			Confusion Matricies	Accuracy
KNN	[[22 8]	0.409836066		KNN	[[23 8]	0.524590164
	[28 3]]				[21 9]]	
R.Forest Test	[[14 16]	0.442622951		R.Forest Test	[[16 15]	0.540983607
	[18 13]]				[13 17]]	
R.Forest Train	[[15 15]	0.459016393		Log. Reg	[[3 4]	0.5625
	[18 13]]				[3 6]]	
R. Forest (10)	[[15 15]	0.508196721		SVM	[[4 3]	0.5625
	[15 16]]				[4 5]	
SVM	[[1 5]	0.5625		R.Forest Train	[[17 14]	0.573770492
	[2 8]]				[12 18]]	
Log. Reg	[[4 2]	**0.6875**		R. Forest (10)	[[17 14]	**0.62295082**
	[3 7]]				[9 21]]	

Our results show that we are able to predict the sidewalk label based on price per square foot with better than 60% accuracy. Higher than random chance, our results do indicate that there is a relationship between fluctuating neighborhood pricing and sidewalk condition.

4.2   Granularity

In conjunction with our regionality results, we wanted to compare it against something more specific. Since we had access to the individual "Project Sidewalk" labels, we worked to grid these points out against the reported crime data. Using random forest modeling, we compared the results from a 25 x 25 grid and a 50 x 50 grid of Washington D.C.
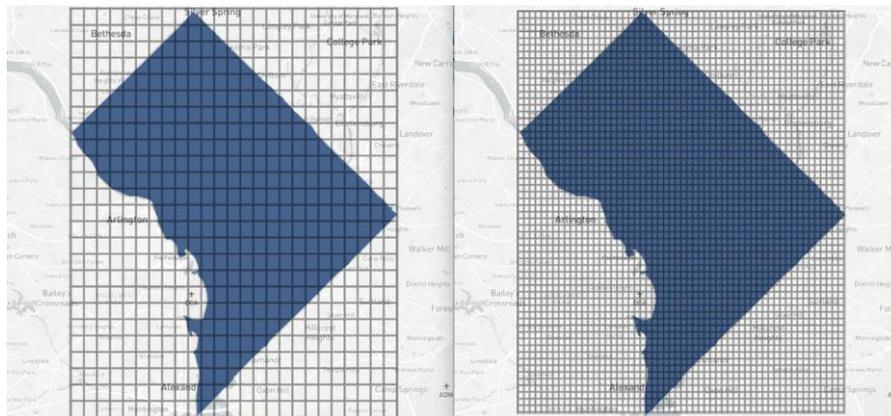

Fig. 3. 25 x 25 Grid Layout of Washington D.C. (*left*), 50 x 50 Grid Layout of Washington D.C. (*right*)

Looking at the different confusion matrices and out of bag (OOB) Error rates, its clear that the 25 x 25 grid provides better results than the 50 x 50 grid at the cost of increased granularity. The "other" label category, assumed to be a catch-all term for all sidewalks that are problematic but do not fit within the "curb ramp" "occlusion" and "surface problem" categories, was not included because the results were inconclusive.

Table 6. Error Calculations for Crime Statistics

		Confusion Matrices		Error	OOB Error
"Occlusion" label					
		0	1	Error	OOB Error
50 x 50		0	1		20.28%
	0	1816	488	0.211805	
	1	19	177	0.09693	
25 x 25		0	1		**14.72%**
	0	412	92	0.1825	
	1	0	121	0	
"No Curb Ramp" label					
		0	1	Error	OOB Error
50 x 50		0	1		6.75%
	0	1733	127	0.06827	
	1	42	598	0.06562	
25 x 25		0	1		**3.36%**
	0	395	13	0.03186	
	1	8	209	0.36866	
"Curb Ramp" label					
		0	1	Error	OOB Error
50 x 50		0	1		4.24%
	0	1749	32	0.01846	
	1	43	693	0.09647	
25 x 25		0	1		**2.88%**
	0	388	3	0.00767	
	1	15	219	0.0641	
"Obstacle" label					
		0	1	Error	OOB Error
50 x 50		0	1		8.76%
	0	1738	182	0.09479	
	1	37	543	0.06379	
25 x 25		0	1		**3.48%**
	0	393	14	0.03439	
	1	10	208	0.04587	
"No Sidewalk" label					
		0	1	Error	OOB Error
50 x 50		0	1		12.00%
	0	1676	201	0.107085	
	1	99	524	0.158908	
25 x 25		0	1		**6.56%**
	0	379	17	0.107085	

	1	24	205	0.158908	

Since the 25 x 25 grid has the lower error rate, we will focus our analysis on the important characteristics that emerged from this grid pattern.

Table 7. Importance Results for Crime Statistics

25 x 25 Grid Importance Results						
	No CurbRamp	No Sidewalk	Obstacle	Occlusion	Surface Problem	CurbRamp	Other
Day	18.4981	18.4294	17.715	17.7626	18.2147	**20.3762**	**17.80**
Evening	**20.8991**	**24.4415**	**21.297**	**25.7356**	**20.9577**	19.6574	17.23
Midnight	14.0016	15.7375	13.6554	14.6961	10.9222	12.0401	15.83
Vehicle Theft	14.6189	9.9354	17.258	9.7994	17.0545	15.6729	14.07
Robbery	12.3292	17.7699	14.5107	17.0238	11.6437	13.6176	12.61
Burglary	11.7326	10.9425	11.6789	11.5311	11.7905	10.0019	10.89
Homicide	1.6579	1.1959	0.1165	1.1552	2.1797	2.2007	4.047
Sex abuse	6.2238	1.4779	3.7681	2.1899	7.2365	6.3536	6.290
arson	0.03837	0.0696	0	0.1059	0	0.0793	1.192

There were several features that stood out amongst the sidewalk data. For all labels, we can see that time of day has the highest level of importance because there is a shift time associated with every entry, however, the distribution of sidewalk labels amongst the entries vary wildly.

Regarding the "No Curb Ramp", "Curb Ramp", "Obstacle", "Surface Problem" and "Other" label, the statistically important crime indicator was vehicle theft. For "No Sidewalk" and "Occlusion", the statistically important crime indicator was robbery. Vehicle theft and robbery are the category leaders, behind time of day, for all sidewalk labels. This could be because these are the most reported crimes for this time period in Washington D.C.

After running statistical analysis on our data, we found that crime is correlated with accessibility data and that housing prices are correlated with accessibility data. More importantly, we can see that crime reported in the "evening" time period stands out as an important factor in determining all the labels. We found that the model with the lowest amount of OOB error was the 25 x 25 grid. This makes sense since our crime data itself was not as granular as the accessibility data. Since much of the reported crime was pending litigation, the location address was generalized to protect the identity of the victims. This contrasts the "Project Sidewalk" data that was granularized down to the exact latitude and longitude coordinates for every label.

Even though our results seem to point to a more regionalized approach, this is beneficial to our application. As a city planner or policy maker, a regional approach will work more efficiently to allocate budget for areas predicted to have problem sidewalks. Using this approach, community leaders can logistically optimize assessors and the crews working on the repairs. Running our 50 x 50 grid, despite producing higher error rates, helped verify our need for regionality.

5   Ethics Awareness

We hope that the methods developed through this paper are used ethically and responsibly. We understand that generating such methods to predict where infrastructure is faulty, opens cities up for lawsuits if users specifically search for inaccessible areas that should be accessible by law. Methods detailed here were developed for the sole purpose of finding a low-cost alternative for identifying and prioritizing problem sidewalks. We appreciate the work of the "Project Sidewalk" team and acknowledge the importance of their efforts to make accessibility accessible.

Additionally, we advocate the use of these datasets in this way, but cannot verify the validity of all crime and housing data. Results could vary based on the collection methods of both of these statistics. We also acknowledge that "Project Sidewalk" consists of labeling completed for and by handicapped persons, not by city planners. Because of this, some of the sidewalks flagged by "Project Sidewalk" may, in fact, be purposefully designed by city planners as a means to divert pedestrian traffic.

6   Conclusion

Using random forest analysis on the Washington D.C. crime statistics and real estate data, we uncovered a predictive relationship for sidewalk accessibility that is better than random chance. The Washington D.C. Department of Transportation reported that even with initiatives in place that prioritize pedestrian mobility, there is still a \$27 million backlog on sidewalk repairs [6]. With the methods we outline here, Washington D.C. will be better able to anticipate sidewalk repairs and better allocate resources to minimize these large backlogs. Additionally, simple repairs can increase the accessibility of the city, resulting in better public health, environmental health, and economic health.

In the spirit of reproducibility, we have utilized publicly available datasets that are updated regularly. While "Project Sidewalk" has only managed to asses the sidewalk of Washington D.C., our methods here have shown that just by using housing prices and crime statistics, we are able to statistically predict problem sidewalks better than random chance. Other cities can reproduce similar results with the local crime and real estate pricing for their area. In this respect, we advocate for increased efficiency by promoting a proactive attitude toward infrastructure repair, instead of the current reactive system. Furthermore, the analysis completed here can be completed at low costs. Currently, the reactive system in place is expensive, whereas, using our methods are low cost and the recommendations can be applied at any budget.

5.1   Future Work

While the findings here help plan for potential sidewalk repairs, moving forward, we would also like to investigate ways to prioritize sidewalk repairs. Next steps include analysis of pedestrian traffic patterns in the Washington D.C. area. Sidewalks near

schools or public transportation should be prioritized higher than those with low pedestrian traffic. We would also like to look into sidewalks at intersections with high amounts of traffic fatalities to determine if sidewalk repairs would help eliminate some of these accidents.

References

1. Boyle, A., Barrilleaux, C. and Scheller, D. (2013). Does Walkability Influence Housing Prices?. Social Science Quarterly, 95(3), pp.852-867.
2. Carr, Lucas. J., Shira I. Dunsiger, and Bess H. Marcus. 2010. Validation of Walk Scores for Estimating Access to Walkable Amenities. British Journal of Sports Medicine 45(14):1144–48.
3. Cesme, Dock, Westrom, Lee, Barrios (2016) Data-Driven Urban Performance Measures: A Case Study Application in the District of Columbia. Submitted for the TRB 96th Annual Meeting for Presentation and Publication.
4. Comp.ddot.dc.gov. (2017). Available at: https://comp.ddot.dc.gov/Documents/Sidewalk%20Curb%20and%20Gutter%20FAQs.pdf [Accessed 7 Sep. 2017].
5. Ddot.dc.gov. (2017). The Vision Zero Initiative | ddot. [online] Available at: https://ddot.dc.gov/page/vision-zero-initiative [Accessed 7 Sep. 2017].
6. Ddot.dc.gov. (2017). Available at: https://ddot.dc.gov/sites/default/files/dc/sites/ddot/publication/attachments/Final%20Operations%20Performance%20Data%20Fourth%20Quarter%20FY%202014.pdf [Accessed 7 Sept. 2017].
7. Hara, K., Le, V., Froehlich, J., (2013). Combining Crowdsourcing and Google Street View To Identify Street-level Accessibility Problems. Proceedings of CHI 2013, Paris, France.
8. Haughwout, A. (2002). Public infrastructure investments, productivity and welfare in fixed geographic areas. Journal of Public Economics, 83(3), pp.405-428.
9. LaCour-Little, Micheal., Melpezzi, Stephen. Gated Community and Property Values. (2001) Presented at Urban Economics Association meeting in Washington, DC.
10. Leinberger, Christopher B., and Mariela Alfonso. 2012. Walk This Way: The Economic Promise of Walkable Places in Metropolitan Washington, DC. Washington, DC: Brookings.
11. Mathur, S. (2008). Impact of Transportation and Other Jurisdictional-Level Infrastructure and Services on Housing Prices. Journal of Urban Planning and Development, 134(1), pp.32-41.
12. Pavlov, A. (2000). Space-Varying Regression Coefficients: A Semi-parametric Approach Applied to Real Estate Markets. Real Estate Economics, 28(2), pp.249-283.
13. Pivo, G. and Fisher, J. (2011). The Walkability Premium in Commercial Real Estate Investments. Real Estate Economics, 39(2), pp.185-219.
14. Roberts, J., Rodkey, L., Ray, R., Knight, B. and Saelens, B. (2017). Electronic media time and sedentary behaviors in children: Findings from the Built Environment and Active Play Study in the Washington DC area. Preventive Medicine Reports, 6, pp.149-156.
15. Rspatial.org. (2017). 5. Spatial distribution models — R Spatial. [online] Available at: http://rspatial.org/analysis/rst/5-global_regression.html [Accessed 7 Sept. 2017].
16. Sidewalk.umiacs.umd.edu. (2017). Project Sidewalk. [online] Available at: https://sidewalk.umiacs.umd.edu [Accessed 7 Sep. 2017].
17. Tappe, K., Glanz, K., Sallis, J., Zhou, C. and Saelens, B. (2017). Children's physical activity and parents' perception of the neighborhood environment: neighborhood impact on kids study.

18. US EPA. (2017). Health Effects of Ozone Pollution | US EPA. [online] Available at: https://www.epa.gov/ozone-pollution/health-effects-ozone-pollution [Accessed 7 Sept. 2017].

19. Wemovedc.org. (2017). [online] Available at: http://www.wemovedc.org/resources/Final/Part%202_Plan_Elements/Pedestrian.pdf [Accessed 7 Sept. 2017].

Appendix

For all code and files referenced within this document, please visit https://github.com/clairecDS/ProjectSidewalk_Remix.