Modeling Urban Capacity with Public Data: Helping Realize Universal Prekindergarten in New York City

Richard Dunks, Ravi Shroff, Jeongki Lim, Miguel Castro, Haozhe Wang
{rad416, rs4606, jl2684, mac1077, hw1067}@nyu.edu
New York University
Center for Urban Science & Progress

July 15, 2014

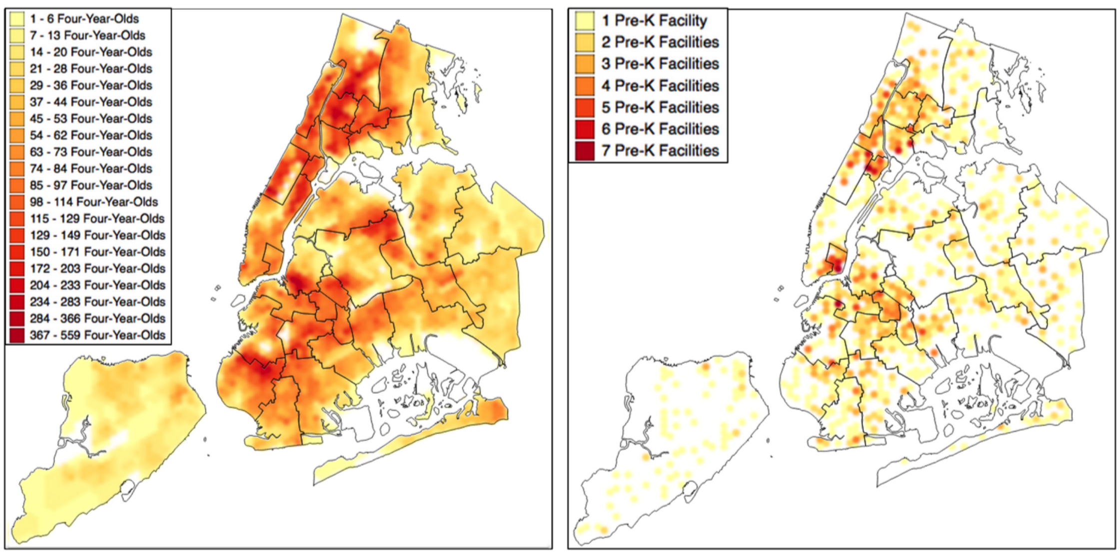

Figure 1: Distribution of four year olds in New York City based on data from US Bureau of the Census (left) and the distribution of publicly funded prekindergarten programs (right)

Abstract

Using public data from the US Bureau of the Census and the New York City Department of Education, we consider the problem of identifying locations in New York City where four year olds are potentially underserved by available capacity in city-funded prekindergarten programs. We were able to create a map of four year old children in New York City and the available prekindergarten capacity in both public schools and publicly-funded community-based organizations. We implement a random allocation algorithm to identify and map underserved locations. As part of initial work in this area, our model incorporates a rough travel distance measurement and has potential for being improved with better travel estimates and a more nuanced modeling of travel flows in New York City, as well as more sophisticated techniques for estimating the location of four year olds in New York City using administrative data on residential tax lots.

1 Introduction

The implementation of a publicly funded universal full-day prekindergarten (pre-K) program has become an important topic in New York City politics after having figured prominently in the 2013 campaign for Mayor of New York City. Studies have shown numerous benefits from high quality pre-K instruction, including increased cognitive abilities, higher test scores in the short term, and access to higher paying jobs in the long term (see Gormley, 2004 for a fuller discussion of these effects).

Currently, public pre-K programs in New York City are offered in public schools and community based organizations (CBOs) receiving public funds. Programs are half day (two hours and 30 minutes of instruction) or full day (six hours and 20 minutes of instruction). While admission to CBO programs is first come, first served, admission to public school programs requires parents to apply by the end of the previous school year and select up to 12 schools for their child to attend. Admission decisions are based on a priority list giving first priority to students who live in the zone for that particular school, then students living in the school district but without a pre-K site in their zone, students living in the same school district but not in the same school zone as the school, students living in the same borough as the school, and finally out-of-borough students, in that order. At each point, having a sibling already attending the school increases the prioritization of an applicant for a seat at that particular school (Department of Education, 2014).

The NYCDOE estimates there are 73,250 children who require access to full-day prekindergarten programs. This number is derived by taking the 81,748 children enrolled in kindergarten and subtracting the estimated 8,498 children who will likely enroll in private pre-K. There are currently 58,528 pre-K seats available in public schools and in publicly-sponsored pre-K programs run by CBOs under contract to either the NYC Department of Education (NYCDOE) or the NYC Administration for Children's Services (NYCACS), which manages pre-K for low-income children with support from the federal government. This includes 26,364 half-day seats and 32,164 full-day seats. Of the currently available seats, 23,671 are in public schools and 34,857 are in CBOs, although the full-day seats are almost evenly distributed between public schools and CBOs (Office of the Mayor, 2014).

To meet the gap between children requiring full time pre-K and the current capacity (approximately 41,000 children), the NYCDOE plans to expand the pre-K system by 23,640 in the 2014-2015 school year, split evenly between the conversion of part time seats to full time seats and an increase in the number of available seats, and 17,446 in the 2015-2016 school year, with most of the increase coming in the conversion of half-day seats to full time seats (Office of the Mayor, 2014).

The purpose of this paper is to use publicly available data to model the distribution of four year olds in New York City and identify areas currently underserved by existing pre-K capacity. Our intention is to develop a more sophisticated approach to modeling capacity that can better inform the decision making process around increasing access to high-quality pre-K, particularly in areas where the added benefit of pre-K instruction can matter most. Our work is preliminary and we note areas where both the available data and methodology could be improved.

2 Data Sources

We use publicly available data from the US Bureau of the Census, the New York City Department of Education, and PediaCities, a platform to "curate, organize, and link data about cities"[1], in order to identify areas currently underserved by existing pre-K capacity. The combination of two open data sets from different governmental organizations (federal and city) present unique challenges to the task of analyzing public data for the public good. We use standard geospatial processing techniques to combine and visualize the data, as well as standard statistical methods for estimating the distribution of four-year-old children throughout the city.

2.1 Census Data

The number and location of four year olds was derived from the 2012 5-year aggregated American Community Survey (ACS). The ACS groups the population into various age bands for reporting purposes. To estimate the number of four year olds living in a particular census tract, the population under the age of 5 (the age band in which the Census reports four year olds) was assumed to be evenly distributed among the 5 years encompassed by children ages 0 - 4. The population number was divided by 5 to arrive at the estimate of four year-olds for a given census tract. This yields an estimate of 105,410 four year olds in New York City[2].

2.2 Prekindergarten Location and Capacity

Data on schools was gathered from both the NYCDOE open data portal[3], and PediaCities. Each pre-K site is identified by a six digit site ID, with the first two digits indicating the school district number, the third digit a letter indicating the borough, and the remaining three digits a unique identifier for the site. From the NYCDOE data we were able to collect capacity information for the public school (PS) sites while PediaCities provided locations for both the PS and CBO sites. For the 2013-2014 school year, NYCDOE lists 1,406 pre-K sites. Of these, 29 public schools listed as having provided pre-K at some point in the past no longer have

[1] Information about PediaCities is available from their website, http://www.pediacities.com/

[2] The disparity between this number and the number used in Office of the Mayor, 2014 is discussed below

[3] Available at http://nycdoe.pediacities.com/

pre-K seats. One school, PS 051 Elias Howe, hasn't had pre-K seats since the 2010-2011 school year, but will have seats for the 2014-2015 school year. To keep the analysis consistent, this school was excluded from this analysis. Among the community-based organizations (CBOs), 6 sites were listed without site IDs and appear to no longer host pre-K seats. These have also been excluded from the analysis.

There are 34,857 CBO seats reported at approximately 850 sites (Office of the Mayor 2014). This matched the 855 CBOs listed in the PediaCities data. Of these 855 CBO sites, NYCDOE had readily available[4] capacity data for only two sites, Baychester Academy and the Staten Island School of Civic Leadership with reported capacity of 36 and 18, respectively. Subtracting these 54 seats from the CBO total leaves 34,803 seats spread across 853 facilities, or an estimated 40.8 seats per facility. For the purposes of this analysis, each CBO for which the actual capacity is unknown has been assigned a capacity of 41.

For the purposes of this analysis, equal-sized hexagonal polygons 650 meters across were used to divide the city. This created a common spatial unit of analysis among the various geographical divisions, allowing for data to be broken down into smaller units and then aggregated as necessary to compare capacity within various political, administrative, and other delineations.

[4]The NYCDOE did not release comprehensive CBO capacity information for the 2013-2014 school year in a machine readable format. An audit of CBO capacity would require going through PDF directory listing intended for parents to select a pre-K to match sites to ensure they appear in the current listing of available programs.

3 Methods and Tools

We divided this task into two main phases. In the first phase, we mapped the the locations of both four year olds in New York City and the pre-K locations. We also added attributes for the pre-K sites based on available information. In the second phase, we applied an allocation algorithm to model the availability of pre-K seats for the population of four year olds we mapped in the first phase.

3.1 Mapping Distribution of Four Year Olds

The US Bureau of the Census releases demographic data aggregated into various statistical subdivisions that follow political and administrative boundaries. The common unit of analysis is the census tract, an area with approximately 1,200 to 8,000 people that can vary in size depending on the population density of a given area. In addition to tracts, the Census releases data in blocks, block groups, zip codes, Census designated places (CDPs), counties, metropolitan and micropolitan statistical areas, states, tribal areas, and the country as a whole. For the purposes of this analysis, data for census tracts were used to estimate the population of four year olds in New York City.

In order to create a common spatial unit for analyzing the population of four year olds and the closely located pre-K capacity over a consistent area, we used a mesh of 2,930 equal-sized hexagons measuring approximately 650 meters across. The size was chosen to provide a unit of analysis that was small enough to provide an estimate of a smaller area than that generally covered by a census tract without being too small that it created a computational task that couldn't have been accomplished on our available hardware. We then employed a technique known as dasymetric mapping to estimate the number of four year olds living in a particular hexagon.

Dasymetric mapping is a technique for disaggregating spatial data into smaller units of analysis based on ancillary information. The technique was first described in 1911 by Benjamin Semenov-Tian-Shansky and employed in his 1923 "Dasymetric Map of European Russia" that used the land use categories to estimate population densities (Petrov, 2012). Dasymetric mapping has since been developed to employ various techniques to match data between spatially mismatched delineations (Mennis, 2009; Zandbergen and Ignizio, 2010).

In this analysis, we employed a basic areal weighting that takes into account the area overlap between the two spatial units, in this case the census tract and the hexagon. Assuming the demographics of a census tract are evenly distributed across the tract, a portion of the demographic characteristics equal to the portion of the tract area covered by the hexagon is then added to the hexagon. For example, if a hexagon completely overlapped the census tract, the hexagon included the entire estimated population of four year olds. If the hexagon only overlapped 50% of the census tract, the hexagon received only half the four year old population. The number of four year olds from each of the constituent census tracts was then summed and rounded to the nearest integer to provide a number of four year olds in the hexagon.

The statistical delineations used by the Bureau of the Census follow administrative boundaries. In the case of New York City, this means that many of the 2,167 census tracts include a significant amount of water area as the boundaries between New York and New Jersey, as well as borders between the 5 boroughs, fall in the middle of local waterways. Assuming that most, if not all people live on land, we excluded the water area from the areal interpolation method described above and used census tracts clipped to the shoreline of New York City[5].

Using the locations of pre-K sites from the NYCDOE available from PediaCities, the number of public school (PS) and CBO pre-K sites were joined to each hexagon. The number of seats were summed for all sites located within a hexagon, separated by whether they were PS seats or CBO seats. No distinction was made between part time and full time seats. Thus a site with 18 morning seats

[5] We used 2010 census tract files prepared and released by the NYC Department of City Planning that exclude the water area from and available on their website at `http://www.nyc.gov/html/dcp/html/bytes/districts_download_metadata.shtml`

and 18 afternoon seats would have a capacity of 36. Each hexagon was also assigned to a particular school district . q For those hexagons lying along district boundaries, the district that occupied 50% or greater of the hexagon area was assigned as the district for that hexagon.

To simulate the ability of parents to take their kids to nearby pre-Ks outside their hexagon, we created a list of nearby hexagons based on a simple distance measurement. A nearby hexagon was defined as a hexagon within 2,000 meters of the given hexagon. This simulates approximately 15 minutes of travel time. We chose this estimate in order to validate the approach with the intention of employing a more sophisticated approach, which we outline below.

Algorithm 1

Input: Hexagons $\mathcal{H} = \{H_n\}_{n=1}^{2930}$,
$H_n = \{P_{H_n}, S_{H_n}^{ps}, S_{H_n}^{cbo}, C_{H_n}^{ps}, C_{H_n}^{cbo}, \mathbf{F}_{H_n}, \mathbf{G}_{H_n}\}$
Initialize parameters:
$P_{H_n} = P_{H_n}^0 \; \{\forall n\}$
$S_{H_n}^{ps} = 0 \; \{\forall n\}$
$S_{H_n}^{cbo} = 0 \; \{\forall n\}$
$\mathcal{X} = [H \in \mathcal{H} \text{ s.t. } H \text{ usable}]$
while length($\mathcal{X}$) $\neq 0$ **do**
randomly choose $H \in \mathcal{X}$
randomly choose non-full neighbor K of H from either $\mathbf{F_H}$ or $\mathbf{G_H}$
if K chosen from $\mathbf{F}_H$ **then**
$S_K^{ps} = S_K^{ps} + 1$
else
$S_K^{cbo} = S_K^{cbo} + 1$
end if
$P_H = P_H - 1$
$\mathcal{X} = [H \in \mathcal{H} \text{ s.t. } H \text{ usable}]$
end while
Output: Pairs $\{(n, o_n)\}_{n=1}^{2930}$
where $o_n = \frac{P_{H_n}}{P_{H_n}^0}$ if $P_{H_n}^0 \neq 0$, and $o_n = -1$ otherwise

3.2 Calculating Need

Our allocation algorithm (Algorithm 1) uses Monte Carlo methods to determine areas of New York City underserved in terms of pre-K program access. The input consists of the geographic distribution of four-year-olds and current PS and CBO pre-K capacities for each of the 2,930 hexagons described above. We simulated the ability of parents to take their children to nearby pre-K seats by creating for each hexagon H a list of "nearby" hexagons, defined to be those hexagons overlapping a disc of fixed radius (the "travel distance") centered at the given hexagon. We assume that although a child in H may attend any CBO pre-K program in any nearby hexagon, he or she may only attend those public school pre-K programs in nearby hexagons in the same school district as H. This assumption is consistent with public school pre-K admissions criteria, which give strong preference to an applicant whose residence is in the same school district as the target school. Without detailed family information, we were unable to model the sibling-preference in PS prioritization.

Each hexagon H has resident population of children P_H, current number of assigned public school and CBO students S_H^{ps} and S_H^{cbo} respectively, public school and CBO capacities C_H^{ps} and C_H^{cbo} respectively, a list of nearby hexagons $\mathbf{F}_H = \{F_1, \ldots, F_{j_H}\}$ in the same school district as H, and a list of all nearby hexagons $\mathbf{G}_H = \{G_1, \ldots, G_{k_H}\}$ (so $\mathbf{F_H} \subseteq \mathbf{G_H}$). Initially, S_H^{ps} and S_H^{cbo} are set to zero and P_H is set to P_H^0, the initial resident population.

We say any hexagon K is a *non-full neighbor* of H if $K \in \mathbf{F}_H$ and $S_K^{ps} < C_K^{ps}$ or $K \in \mathbf{G}_H$ and $S_K^{cbo} < C_K^{cbo}$. We say H is *usable* if it has at least one child ($P_H > 0$), and at least one non-full neighbor.

The algorithm then works as follows. While there exists at least one usable hexagon, randomly choose such a usable H and randomly choose a non-full nearby hexagon K from either $\mathbf{F_H}$ or $\mathbf{G_H}$. Decrement Hs resident population P_H by 1 and if K was chosen from $\mathbf{F_H}$, increment Ks assigned public school students S_K^{ps} by 1, otherwise increment K's assigned CBO students S_K^{cbo} by 1.

When the algorithm terminates, for each hexagon H compute the output statistic P_H/P_H^0, the percentage of resident children in H that were unable to be allocated to a pre-K spot. We chose to run the algorithm six times due to computational limitations and average the output statistics. Finally, we visualized the averaged output statistic by creating a map where each hexagon was shaded from green to red based on the output statistic (red means all children were unallocated and green means all children were allocated to pre-K seats). Hexagons that did not possess children to begin with are shaded white. For work related to this approach, see Holmberg, et al., 1999.

3.3 Software Tools

We parsed the available text files using Python scripts before uploading them to a PostgreSQL database hosted on the Amazon Web Services cloud computing platform. We used the PostGIS spatial extension to perform the areal weighting of census tracts to hexagons and the dasymetric spatial join of demographic data to hexagons. We coded the simulation algorithm in Python to run on data exported from the PostgreSQL database, joining results back to the database at the conclusion of the program. We visualized the data using QGIS, an open-source geospatial information system (GIS). All tools used in this analysis are open-source and freely available online.

4 Results and Discussion

Our analysis shown in Figure 2 identified areas in all 5 boroughs, particularly the neighborhoods in Sunset Park in Brooklyn (red box lower left), Corona (middle right) and Far Rockaway (bottom right) in Queens, the Upper East Side and the Upper West Side in Manhattan (middle left), and the North Bronx (top right). In these neighborhoods, we estimate there is a high percentage of four year olds who aren't able to find a pre-K slot nearby based on 2013-2014 capacity.

Figure 3 shows a similar analysis looking at the estimated number of four year olds unable to find a spot (as opposed to the percentages shown in Figure 2). While the areas

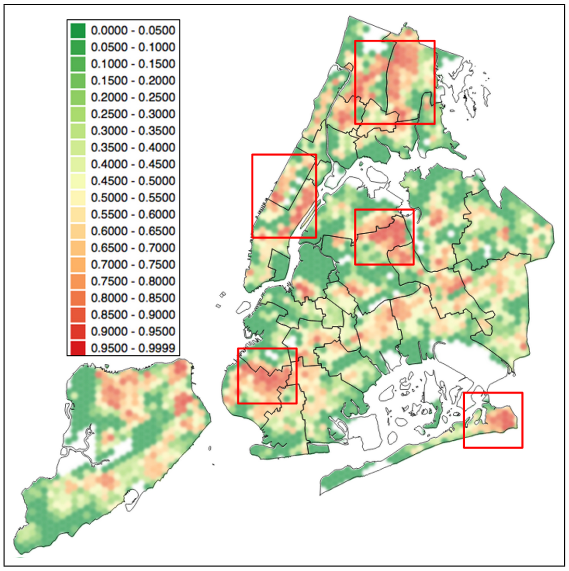

Figure 2: Percentage of four year olds in a given hexagon that are unable, on average, to find a pre-K slot, with highly underserved areas noted (red boxes). NYC Department of Education school district boundaries are noted (black lines).

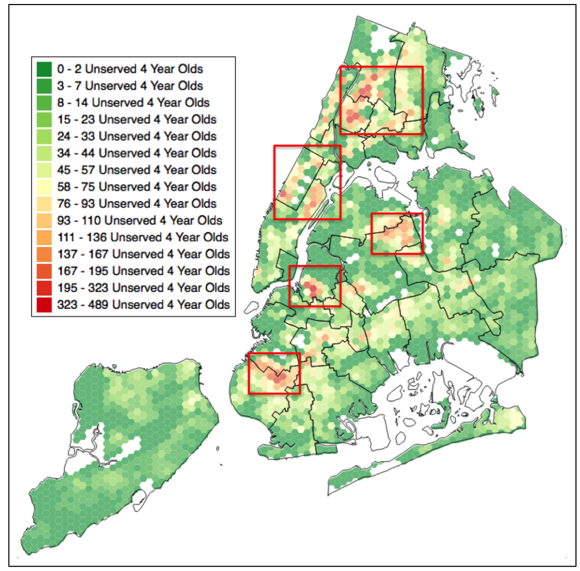

Figure 3: Raw numbers of four year olds in a given hexagon that are unable, on average, to find a pre-K slot, with highly underserved areas noted (red boxes). NYC Department of Education school district boundaries are noted (black lines).

of Sunset Park in Brooklyn and Corona in Queens are underserved by percentage, Sunset Park shows a higher number of four year olds who likely won't be able to find a pre-K spot nearby. The contrast between Sunset Park and Manhattan areas is interesting as Sunset Park is known as a low-income area with a number of immigrants while the underserved areas of Manhattan are known as high-income areas where parents likely have access to private schools rather than rely on publicly funded options. This suggests that impediments to truly universal prekindergarten exist at both ends of the socio-economic spectrum.

5 Limitations and Future Work

Ultimately, the reliability of this approach rests on the accuracy and completeness of the data upon which it is built. Not having accurate CBO capacity information prevents us from making an accurate estimation of the available capacity throughout New York City. Having this data easily available in a machine readable format would not only enable work such as ours, but also encourage entrepreneurs to develop web and mobile applications to provide parents with this critical information.

Our population estimate of 105,410 estimates all four year olds in New York City whether they are enrolled in public school or not. The figure of 81,748 four year olds used in Office of the Mayor (2014) is based on the number of five year olds enrolled in public kindergarten. The difference of 23,662 is possibly explained by the number of five year olds enrolled in private school; however the 2012 5-year ACS estimates this number to be 10,855. It's unclear where this discrepancy comes from and suggests a considerable difference in how the NYCDOE and US Bureau of the Census estimate population in New York City.

Our use of a simple dasymetric modeling approach with areal weighting likely doesn't provide the most accurate estimate of four year olds for a given area. This could be refined by using administrative data on the location of tax lots with residential units similar to work in Maantay, et al. 2007 and Maantay, et al. 2008, which used data from the NYC Primary Land Use Tax Lot Output (PLUTO) database to more accurately model the impact of limited access highways in New York City on asthma rates in the Bronx. PLUTO provides information on the various tax lots around NYC, including the number of residential units on a particular tax lot. For example, a tax lot with 10 units would thus have roughly double the number of occupants as a tax lot with 5 residential units. The fraction of the total population of the census tract residing on that particular tax lot would be determined by the number of residential units on that particular tax lot divided by the total number of residential units in the entire tract.

Assuming the two tax lots mentioned above were the only buildings in a particular census tract, the total number of residential units in the tract would be 15, with the first tax lot of 10 residential units having 2/3rds of the population and the second tax lot of 5 residential units having 1/3rd of the population. In the case of four year olds, if the population of four year olds was estimated to be 9, the first tax lot (of 10 residential units) would be estimated to have six four year olds while the second tax lot (with 5 units) would be assumed to have three four year olds. With this information, it would be possible to more accurately model the number of four year olds within the census tract instead of just assuming an even distribution across the entire tract where there could be large areas of non-residential space.

Accounting for public transit options that allow distant areas to be easily accessed in a reasonable amount of time would help provide a better estimate of the capacity available to a child at any given location. Incorporating travel by public transit would increase the number of hexagons to which a family could travel and increase the complexity of the calculation but more accurately estimate the real-world choices of parents.

Modeling the predominant direction of that flow would also help create a more reliable estimate of access, as the flow of New Yorkers tend to flow towards the business districts in lower and midtown Manhattan, downtown Brooklyn, and Long Island City in Queens, and back to the outer boroughs in the evening, rather than flowing in all available directions equally at all times. Census journey to work data available through the Census Transportation Planning Products (CTPP) Program [6]could help refine the analysis and give a better estimate of where parents would like to take their children.

Beyond these refinements to the methodology, having access to the data available in public records, such as birth, immunization, and public assistance records, could greatly improve the modeling techniques we employ, producing a much more accurate map of where four year olds live in New York City. As this is legally protected data, such work would need to be carried out under procedures outlined in federal, state, and local law, but could yield a highly accurate map of the four year old population in New York City.

6 Conclusion

We've outlined an approach that uses open data and open-source technology to reasonably estimate the distribution of four year olds in New York City using Census data and the available capacity in publicly funded pre-K programs using data from the New York City Department of Education. This approach gives decision makers the ability assess need in a quantitative way. Using additional demographic information, it becomes relatively easy to identify where that need exists, whether in low income areas, such as Sunset Park or higher income areas like the Upper East Side of Manhattan. Given limited resources, potential sites could be evaluated based on the likely impact on the local community given a set of criteria for evaluating the relative impact of one site over another, with the goal of providing capacity to those it's likely to benefit most.

Our work is preliminary, with a number of methodological enhancements and improvements in data sources noted, but there are implications for how demand for school seats is modeled for all grades, providing a more sophisticated, data-driven approach to optimizing capacity to meet demand. We hope to make an interactive tool based on the model we've built in order to allow policymakers and government employees to adjust plans for new capacity in an easy and intuitive way.

The City of New York has embarked on an ambitious expansion of the pre-K program, the success of which will rely greatly on how well available technology is leveraged to achieve program success. To make the system truly universal, a seat should be available for each four year old at or near where they live. The realities of public finance and limitations of infrastructure and resources often mitigate against providing every public service at once and at the level expected. In that case, decision makers must target resources where they are needed most to alleviate the most critical need.

[6]Accessible at `http://ctpp.transportation.org/Pages/default.aspx`

7 Acknowledgements

The authors would like to acknowledge Dr. Sharad Goel, Dr. Manuela Veloso, and Dr. Theodoros Damoulas for their assistance and stimulating conversations regarding this work, as well as the support of Dr. Constantine Kontokosta who supervised this research as part of the Urban Science Intensive I course at the Center for Urban Science and Progress at New York University.

References

[1] City of New York, Department of Education (2014). *Pre-kindergarten Directory, 2014-2015.* Retrieved from `http://schools.nyc.gov/NR/rdonlyres/AAEE65E0-8326-4E33-BACA-A462A2ECC65E/0/201415PreKDirectoryUPDATED07012014.pdf`

[2] City of New York, Office of the Mayor (2014, January). *Ready to Launch: New York City's Implementation Plan for Free, High-Quality, Full-Day Universal pre-Kindergarten.* Retrieved from `http://www1.nyc.gov/assets/home/downloads/pdf/reports/2014/Ready-to-Launch-NYCs-Implementation-Plan-for-Free-High-Quality-Full-Day-Universal-pre-Kindergarten.pdf`

[3] Eicher, C. L., and Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: implementation and evaluation. *Cartography and Geographic Information Science*, 28(2), 125-138.

[4] Gormley, W., Gayer, T., Phillips, D., and Dawson, B., 2004. The Effects of Universal Pre-K on Cognitive Development. Washington, DC: Georgetown University, Center for Research on Children in the U.S.

[5] Holmberg, K.; Rönnqvist, M.; and Yuan, D. (1999). An exact algorithm for the capacitated facility location problems with single sourcing. *European Journal of Operational Research*, 113(3), 544-559.

[6] Maantay, J. A., Maroko, A. R., and Herrmann, C. (2007). Mapping population distribution in the urban environment: The cadastral-based expert dasymetric system (CEDS). *Cartography and Geographic Information Science*, 34(2), 77-102.

[7] Maantay, J. A., Maroko, A. R., and Porter-Morgan, H. (2008). Research note - A new method for mapping population and understanding the spatial dynamics of disease in urban areas: Asthma in the Bronx, New York. *Urban Geography*, 29(7), 724-738.

[8] Mennis, J. (2009). Dasymetric mapping for estimating population in small areas. *Geography Compass*, 3(2), 727-745.

[9] Petrov, A. (2012). One hundred years of dasymetric mapping: back to the origin. *The Cartographic Journal*, 49(3), 256-264.

[10] Zandbergen, P. A. (2011). Dasymetric mapping using high resolution address point datasets. *Transactions in GIS*, 15(s1), 5-27.

[11] Zandbergen, P. A., and Ignizio, D. A. (2010). Comparison of dasymetric mapping techniques for small-area population estimates. *Cartography and Geographic Information Science*, 37(3), 199-214.