Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Estimating small area demand for online package delivery

Tayo Fabusuyi\textsuperscript{a,}\textsuperscript{*}, Richard Twumasi-Boakye\textsuperscript{b}, Andrea Broaddus\textsuperscript{b}, James Fishelson\textsuperscript{c}, Robert Cornelius Hampshire\textsuperscript{d}

\textsuperscript{a} Transportation Research Institute, University of Michigan, 2901 Baxter Rd, Ann Arbor, MI, United States of America
\textsuperscript{b} Robotics and Mobility Research, Ford Motor Company, United States of America
\textsuperscript{c} tOS Studio, Robotics and Mobility Research, Ford Motor Company, United States of America
\textsuperscript{d} Gerald R. Ford School of Public Policy, Transportation Research Institute, University of Michigan, Ann Arbor, MI, United States of America

ARTICLE INFO

Keywords:
Online delivery
Package delivery
Small area estimation
National household travel survey
Synthetic population
Miami-Dade county

ABSTRACT

Using publicly available microdata sets, we show how estimates for online delivery purchases can be generated for small geographic areas defined in our study as micro analysis zones (MAZ) and how these estimates vary across the MAZs that featured in our study. With a focus on Miami-Dade County, we use both the national household travel survey (NHTS) data and synthetic data obtained from Southeast Florida Regional Planning Model (SERPM) to generate demand estimates of online delivery purchases for more than 5300 distinct geographic units in Miami-Dade County. We assess the quality of the estimates using measures of predictive accuracy and by comparing the cumulative values obtained with the population estimates generated from the NHTS survey data for Miami-Dade County. Our approach fills a void in the area of purchases of online delivery items where rich observable data are typically unavailable and it also provides the added potential benefit of being easily replicated nationwide given the emphasis on the use of publicly available data.

1. Introduction and background context

According to the 2017 National Household Travel Survey (NHTS), over 55% of people made at least one online purchase in the last month that required home delivery. This is a 12% increase over the 2009 NHTS figure.\textsuperscript{1} This signals a rapid rise in urban goods delivery due to online shopping and app-based on-demand delivery services. In 2017, the market for goods delivery was estimated at $12.5 billion (Joerss et al., 2016), and this is expected to double over the next 10 years. This growth is driven in part by the benefits that accrue to consumers including convenience and access to a variety of products at the touch of a button. However, the societal costs of this convenience have not been evaluated. While online ordering and delivery reduces consumer travel time and the associated traffic congestion costs to shop and pick up items, on the flip side it generates its own set of traffic congestion costs and adds significant packaging costs for the home deliveries.

At present, many cities are searching for ways to decrease the environmental impacts of urban goods delivery. Some of these measures are requiring off-peak delivery times using small vehicles, as well as active curbside management. In addition, private actors are introducing innovative technology-driven delivery systems including on-demand delivery services, crowd-sourced delivery services, pick-up and drop-off lockers, drones, and driverless vehicles (Cardenas et al., 2017; Lachapelle et al., 2018). However, at present, there are few if any public sources of data about goods delivery at fine geographic scales. Much of the data on the volume and locations of the pickup and delivery of packages in cities is the property of private firms, and yet this data is critical for the construction of more efficient and cost-effective delivery systems. Firms seeking to enter the urban goods delivery market do not have access to microdata about goods delivery, and thus must estimate the demand before entering the market.

The dearth of such data also leaves municipal policy makers, who need to play a role in the coordination of these delivery systems, in the dark. Without accurate data, they are unable to make informed policy decisions in conjunction with the private sector to improve societal benefit. There is however a way this shortcoming could be alleviated by employing NHTS survey data. The survey data has information on the number of times an individual purchases online items or goods for delivery. However, the most granular level of data provided by the NHTS is at the Metropolitan Statistical Area, or MSA level, which limits the utility of the data for efficient package routing purposes. The current effort addresses this shortcoming by presenting two methods to

\textsuperscript{*} Corresponding author.

E-mail addresses: fabusuyi@umich.edu (T. Fabusuyi), rtwumasi@ford.com (R. Twumasi-Boakye), abroaddu@ford.com (A. Broaddus), jfishels@ford.com (J. Fishelson), hamp@umich.edu (R.C. Hampshire).

\textsuperscript{1} FHWA NHTS Brief: Changes in Online Shopping Trends, August 2018, https://nhts.ornl.gov/assets/NHTSBriefOnlineShopping081018.pdf

https://doi.org/10.1016/j.jtrangeo.2020.102864

Received 9 March 2020; Received in revised form 13 September 2020; Accepted 17 September 2020
Available online 22 September 2020

0966-6923/ © 2020 Elsevier Ltd. All rights reserved.
generate small area estimates of the demand for goods delivery, using as a starting point the 2017 NHTS data coupled with synthetic household and individual records obtained from the South East Florida Regional Planning Model (SERPM). The SERPM is an activity-based travel demand model adopted for South East Florida with the objective of providing travel demand forecasts. More information is available at https://www.fasuonline.net/index.php?/model_pages/modD44/index/.

Estimates of the frequency of online purchases for at-home delivery reveal insights on how cities can better utilize their parking facilities or manage their curbside spaces. For example, cities can better monitor how curbside spaces are used including determining activities that should have priority for specific times. Even more impactful is the temporal dimension of these purchases including the product type – food/meal, fresh grocery, or packages. Having this information enables a city to coordinate across the disparate set of activities that use curbside spaces and be able to dynamically manage them instead of the static approach that is the norm in several cities. The information could be used in making policy changes such as curbside access rules and for pricing decisions with the objective of improving safety and creating more efficient systems that meet the demands of residents.

The majority of the previous work on estimating the demand for goods delivery has utilized surveys from private market research firms (Visser et al., 2014; Allen et al., 2019). The methods have included online interviews, focus groups, and analysis of privately held industry data. Unfortunately, disaggregated data are not publicly available from these approaches. Some of these studies have distinct population segments or geographical focus. For example, a recent study by Sousa et al. (2020) reported online shopping disparities between rural and urban consumers, and another effort by Unnikrishnan and Figliozzi (2020) studied the impacts of COVID-19 on home deliveries using a survey - they reported that higher income households, younger residents, and those with high technology usage are more likely to use e-commerce. Punel et al. (2018) also highlighted young men and individuals with fulltime employment were more likely to patronize crowd-shipping. Studies with a city focus include those of those of Cao, 2009; Cao et al., 2010; Cao, 2012 and Rahman et al., 2018 study in Dhaka City, Bangladesh which showed consumers’ choice for online shopping is largely to save time and have access to more product options.

Nguyen et al. (2019) used cluster analysis to identify three consumer types that showed preferences for online purchases. These segments were classified as price-oriented, time-and-convenience-oriented, and value-for-money-oriented. Rosa and Yunita (2020) in their survey of respondents who use online applications like GoJek, Grab, McDonalds, KFC, and Pizza Hut, concluded that the ease and preference for online payments may drive increases in online food deliveries. Van Droogenbroeck and Van Hove (2017) studied the adoption of online grocery shopping in the Belgian supermarket chain, Colruyt. They found that consumers’ choice relied more on the household level than the personal level - for example, age captures both a person’s ability and household need for technology as there exist correlations between children and the household working situation.

With a focus on drivers of continued usage of online shopping, a survey-based study in India identified that perceived usefulness of website and perceived risk of online transactions were the most prominent indicators of actual purchase behavior (Singh and Srivastava, 2018). Similar observations were reported by Yeo et al. (2017) on online food delivery. In addition, several works based on the 2009 NHTS survey data analyzed online shopping behavior and deliveries (Zhou and Wang, 2014; Wang and Zhou, 2015). Of these two, the work of Wang and Zhou (2015) is most closely related to this paper. Using the 2009 NHTS, they estimated delivery demand to households in the month before the survey date using a negative binomial regression. The covariates included a range of land use and socio-economic variables. They applied the resulting regression coefficients to a case study in the state of New York. The approach we present in this paper improves on Wang and Zhou along two dimensions. First, we implement a multiple imputation approach (White et al., 2011) to online delivery demand estimation. Second, we enhance the negative binomial approach by coupling it with an empirical Bayes (EB) approach (Carlin and Louis, 2000). This enhances the prediction by combining the national baseline trends with NHTS survey records for a given local area.

The balance of our paper is organized as follows. Section 2 provides information on the geographical area of focus and the sources of the data analyzed in the study while the methodology is discussed in Section 3. Section 4 presents the results including a discussion of the findings. Section 5 concludes and provides insights on areas for further studies.

2. Geographical area and data sources

Miami-Dade County is the geographic area of this study. This southeast Florida county’s 2010 population was approximately 2.515 million individuals and 908,000 households. The county consists of 5345 MAZs; the geographic unit of analysis for the study. MAZs recognize the ubiquity of walking as a travel mode in urban areas, and network assignments on walk travel mode are carried out from one MAZ to another. The spatial boundary of a MAZ varies depending on population and employment figures – ranging from large geographical areas in less densely populated areas such as the suburbs and the rural areas to the size of a city block or less in densely populated areas as shown in Fig. 1. The number of households exhibits a similar pattern. Absolute household numbers per inhabited MAZ ranges from 1 to 3220.

Three different micro datasets are used to carry out the analysis – synthetic household and population data obtained from the SERPM; survey data for both person and household data from the 2017 NHTS survey; and the US Census Public Use Microdata Sample (PUMS) dataset. The first two datasets are the primary datasets used for the analysis, while the PUMS dataset provides the basis on which we link MAZs

![](https://www.fasuonline.net/index.php?/model_pages/modD44/index/)

**Fig. 1.** MAZs’ population density per acre (~0.004 km²) for Miami-Dade County. Source: Authors’ calculation based on SERPM data.
to distinct public use microdata area (PUMA). Merging of both the household and person datasets, either for the SERPM synthetic data or the NHTS survey data is achieved by using the unique household identification numbers. In contrast to the synthetic data, the NHTS survey data are weighted datasets, obtained from a stratified sampling approach with a total number of nationwide population and household records of 264,234 persons and 129,696 households respectively.

Persons and household variables of interest that are common to both the SERPM synthetic data and the NHTS survey data include a dummy variable that captures how densely populated the respondent’s MAZ is. This variable is the only spatially relevant variable featured in our analysis. Other variables specific to individuals include age, employment status, gender and level of education. The balance of the variables is at the household level and include household size, household level of income and the number of vehicles in the household. Coding of the dummies or categorical variables are influenced by the threshold specified from the NHTS codebook. In addition, the choice of these thresholds reflects existing policy measures. For example, the $100,000 threshold is marginally higher than the >400% of the Federal Poverty Level (FPL) income group for a family of four in Miami Dade County in 2017 that the government uses as a cutoff for determining eligibility for the Affordable Care Act (ACA) subsidies. The dependent variable, deliver is a count variable representing the number of times online purchases were made in the last 30 days by an individual and is obtained from NHTS data. Excluding deliver, each set of variables is from both the NHTS survey data and the SERPM synthetic data with the variables harmonized across the datasets using equivalent measures and the same thresholds when creating dummies or categorical variables. The table for these variables including the descriptions is provided in Table 1 below.

3. Methodology

We estimate demand for online purchases at the level of individuals. To simplify our analysis, we assume that the online purchases are delivered at the individual’s place of residence and no distinction is made between families and households. In addition, for practical considerations, we aggregate demand for online purchases for each MAZ and assume that the demand was fulfilled at the centroid of the MAZ.

Generating online demand estimates for individuals is carried out using two approaches – a parametric (count) regression with the estimates modified subsequently using empirical Bayes method, and multiple imputation using chained equations (MICE) algorithm. Both approaches are validated using goodness of fit measures and compared to the ground truth – NHTS online purchase for at-home delivery data for the Miami-Dade County area to establish which of the two approaches provides the better fit. Given that our interest is in obtaining estimates of online purchases for at-home delivery, we examine more than one approach to establish which one provides the best fit relative to the observed value.

To generate the online demand estimates, we require information on the attributes of individuals and the households they are associated with for each MAZ. For us to generate the demand for each MAZ, we need to know not just the population or the number of households for each MAZ but also the attributes of these individuals and households. The NHTS data, however, does not have information specific to individuals or households at a granular geographical level. To work around this limitation, we use synthetic household and persons data generated from SERPM. The synthetic data is made up of households and individuals with the associated list of characteristics such as population density of the MAZ in which the individual lives, household size, number of vehicles, and income, all at the household level, as well as age, gender, employment status and education attainment at the individual level. Due to the limited number of variables covered in the synthetic data from the SERPM, we are constrained by the variables that feature in our analysis. Variables such as access to and frequency of internet use, race, and household lifecycle, that have proven to have good explanatory powers, do not feature in our analysis.

3.1. Combining the data sources

We begin the data combination process by merging both the household and the person records separately for the synthetic SERPM data and the NHTS survey data. The household to individual merging is a 1:m merging using the household ID. After the merging, we identify the MAZs that belong to Miami-Dade County. Of the 12,022 MAZs in the state of Florida, 5345 are in the Miami-Dade County. Subsets of the county can be analyzed in greater detail by narrowing down the MAZs to be targeted using the PUMA these MAZs belong to though this was not carried out in the present analysis. A mirror image of the process of focusing on Miami-Dade County is carried out on the NHTS data at the MSA level. We then link the NHTS online demand purchases data with the synthetic SERPM data using the joining variables common to both the synthetic and survey data. The combined dataset integrates elements of the data sources using the joining variables so that the online demand purchases data and the vector of the overlapping variables are now present in the combined dataset. We ensure that the joining variables are placed on an even platform by standardizing them for variables with different units and creating dummies or categorical variables with the same threshold levels across the two data sources.

3.2. Preliminary analysis on estimating online demand purchases

Given that we are using a parametric approach for estimating the demand for online purchases, we begin by fitting the data to ascertain the distribution that provides the best fit using information criteria measures. The variable of interest is deliver, the count variable that represents the number of times items were purchased online for delivery in the past month. Analysis of the data is limited to those who provided responses that are either zero or positive. Responses for the variable are top coded at 99. This reduces the number of the observations used in the regressions to 228,448. Responses for the deliver variable are linked to the person records.

We fit the “deliver” data into a distribution, selecting the best distribution using a series of information criteria. The negative binomial distribution provides the best fit irrespective of the information criteria (AIC, BIC or Chi-Sq. Statistics) used to evaluate the fit. Compared to the Negative Binomial, the Poisson provides a less stellar fit for the data. Fig. 2 shows the cumulative mass function of the data for the best distribution fit. The summary statistics based on the first and second order moments also confirm this.

The mean of the observations is 2.62 while the variance is 19.36, a figure more than seven times the mean. The Poisson distribution assumes that the first and second order moments are equal, which is not the case here given the summary statistics. The over-dispersed nature of the data with more than 43% of respondents having zero online purchases in the past month is a factor leading to a variance significantly large compared to the mean. The subsequent section elaborates more on this by assuming that the zeros are a combination of two data generating processes - one where some of the respondents have a 100% probability of a zero purchase and others, a number between the range of 0 and 99 drawn from the negative binomial distribution.

---

2 http://aidsnet.org/wp-content/uploads/2016/03/6-2017-Federal-Poverty-Guidelines-ACA-ranges.pdf

3 The household life-cycle measures are derived from the ages, the relationship and the work status of individuals within a household.
3.3. Online delivery demand estimation methods

Given the interest in obtaining estimates for online purchases for at-home delivery, the objective of the regression and imputation method is prediction. Consequently, there is an interest in exploring more than one approach and selecting the one that provides the best fit relative to the observed value. We use two different approaches in generating the online purchase demand estimates thus:

- Having shown that the negative binomial distribution provides the best fit, we start by running a negative binomial regression model to generate the predicted online demand using conditional expectations and then take an empirical Bayes approach. This approach avails us the opportunity to generate data dependent priors obtained using maximum likelihood methods.
- We proceed with the assumption that the on-demand delivery data is missing from the synthetic data. Using the combined dataset, we restrict the missing observations to a random sample of the synthetic data with the analysis repeated without replacement. This approach ensures that the totality of the information in the dataset is utilized instead of being constrained to the variables with the missing data.

The following subsections provide more information on these approaches.

### 3.3.1. Negative binomial regression model (NBRM)

We earlier made the case that the Poisson distribution did not provide a good fit – an indication that the distribution could not sufficiently accommodate the overdispersion in the data. The NBRM addresses this by introducing an extra parameter that captures the unobserved heterogeneity among the observations. The functional form of the NBRM is specified thus:

\[
y_i = \exp \left( \sum_{m=1}^{M} \beta_m X_{im} + \epsilon_i \right) = \exp \left( \sum_{m=1}^{M} \beta_m X_{im} \right) \exp(\epsilon_i)
\]

where \( X_{ij} = 1 \quad \forall i \text{ and } j \), the added error term. The conditional expectations of the demand for online purchases is \( \sum_{m=1}^{M} \beta_m X_{im} \) where the \( (\beta_m) \)s are the coefficient estimates of the negative binomial regression model. \( \exp(\epsilon_i) \), the exponential form of the error term is assumed drawn from a gamma distribution. This assumption lends itself to generating data dependent priors that form the basis of the subsequent empirical Bayes method. The gamma distribution provides the equation that explicitly ties the probability of the draws with \( \alpha \), the parameter that represents the degree of dispersion in the predictions.

We restrict data for the NBRM to the nationwide NHTS data excluding Florida. Ignoring the subscript \( i \), we denote the conditional expected demand for online purchases obtained from the regression model \( \mu_{\text{pred}} \) and the data for Florida, \( y_{\text{obs}} \). The EB estimate makes use of both the observed number of the demand for online purchases (Florida’s NHTS data on purchases made online) and the NBRM predicted number of purchases. The EB estimates of the number of times online purchases were made in the past month for each individual as:

\[
D = w \mu_{\text{pred}} + (1 - w) y_{\text{obs}}
\]

where \( w \) is defined thus:

\[
w = \left(1 + \frac{\mu_{\text{pred}}}{\alpha}\right)^{-1}
\]

| Variable | Description |
|----------|-------------|
| Deliver  | Is a count variable representing the number of times online purchases were made in the last 30 days by an individual |
| Density  | Dummy variable which equals 1 if the MAZ in which the individual is resident has at least a population density of 4000 per square mile and 0, otherwise |
| Prod age | Dummy variable which equals 1 if the individual is in the 18 to 64 age band and 0, otherwise |
| Gender   | Dummy variable which equals 1 if the respondent is female and 0, otherwise |
| Employed | Employment dummy that equals 1 if the respondent is employed, and 0 otherwise. |
| hhcount  | Count variable that represents the number of people living in a household |
| Grad     | An education attainment dummy that equals 1 if the individual has at least a bachelor’s degree, and 0 otherwise |
| Income   | Dummy variable that equals 1 for individuals from households with incomes in excess of $100,000 and 0 otherwise |
| Vehicles | Count variable that represents the number of vehicles in a household |

**Table 1** Description of the variables.
and $\alpha$ is the overdispersion parameter, and the value of $w$ lies between 0 and 1. The regression coefficients and the overdispersion parameter $\alpha$ are estimated jointly using maximum likelihood estimation (Cameron and Trivedi, 2013). The demands are reported both at the level of households and also aggregated to the MAZ level.

### 3.3.2. Imputation method

Our analysis assumes that the delivery demand is missing from the synthetic data. We subsequently address the issue using a multiple imputation method advocated by Rubin (1987). The method generates multiple copies of the dataset and the approach ensures that the totality of the information in the dataset is utilized instead of being constrained to the variables with the missing data. In order to keep a healthy ratio of missing data relative to the data required to generate the imputed data, we randomly choose individuals within the synthetic data and rerun the imputation process multiple times.

Borrowing from Rubin (1987), the estimated variance $T$ of a point estimate $Q$, based on $M$ data imputations is:

$$T = W + (1 + M^{-1})B$$

where $W$ is the within-imputation variance component or the average of the $M$ imputed variances:

$$W = \frac{1}{M} \sum_{m=1}^{M} W_m$$

and $B$, the between imputation variance equals:

$$B = \frac{1}{M - 1} \sum_{m=1}^{M} (\hat{Q}_m - \bar{Q})^2$$

$\hat{Q}_m$ is a point estimate obtained as a coefficient of the regression of the $m$th imputed dataset. The overall variance is made up of two parts: $W$, that captures the variation within each imputed dataset irrespective of the missing data points and a second component $B$ that measures how the estimates vary across the imputed datasets. $(\hat{Q}_m - \bar{Q})$ is the difference between the point estimate of the $m$ dataset and the arithmetic mean of the $M$ complete data estimates with the statistic $T^{-\frac{1}{2}}(\hat{Q}_m - \bar{Q})$ following a $t$-distribution. The imputed data is generated using the multiple imputation by chained equations (MICE) approach that is particularly suited for non-continuous variables (van Buuren, 2007; Lee and Carlin, 2010).

### 4. Results and discussion of findings

We run the NBRM using the 2017 NHTS sample dataset with deliver, the count variable for the number of online purchases made in the last 30 days top-coded at 99 as the dependent variable. The dependent variable is regressed on the explanatory variables that are related either to individuals or the households as defined in Table 1. Given that the NHTS data is a weighted sample dataset, the regression explicitly corrects for this by reflecting the person weights in the analysis. It also corrects for the possible correlation among household individuals by defining household as the cluster unit from which individuals are sampled.

We investigate the presence of multicollinearity by estimating correlation measures among the explanatory variables and by running the coefficient matrix post regression. The highest correlation figure of 0.507 is obtained between productivity age and employment with an associated coefficient matrix figure of 0.460 between the two variables. Variance inflation figures (VIF) range from a low of 1.02 for gender to 1.50 for productivity age with a mean value of 1.30. While we acknowledge that there may be efficiency losses as in higher variances, these figures are appreciably lower compared to suggested tolerable limits (Hair et al., 2014) and thus the fit or the coefficient estimates for each regressor should not be affected. Table 2 presents the regression results for the NBRM. The table reports the coefficient estimates of the NBRM and their associated standard errors.

#### Table 2

| Deliver | Coef. Estimates |
|---------|----------------|
| Density | 0.054*** |
| Productivity age | 0.505*** |
| Gender | 0.248*** |
| Employed | 0.235*** |
| Household size | -0.040*** |
| College graduate | 0.516*** |
| Household income | 0.328*** |
| Household vehicles | 0.058*** |
| Alpha | 1.871 |
| F-test (overall significance) | 404.04 |
| Observations | 226,009 |

*Significant at the 0.10 level; ** at the 0.05 level and *** at the 0.01 level.

The F-test, that provides a measure of the model’s overall fit, shows conclusively that the model is valid given that we can reject the null hypothesis of all the coefficient estimates being equal to zero with extreme high confidence − above 99.9%. In addition, all of the explanatory variables except for density, are significant at the 0.01 significance level. Apart from the linear specification for the size of the household, we model the hhcount variable using a quadratic specification and the relationship is not significant. Density, the density of MAZs was initially modeled as a categorical variable − low for all MAZs having less than 1000 persons per square mile; medium for MAZs with between 1000 and 3999 persons per square mile and high, for MAZs with 4000 or more persons per square mile. The variable is later re-classified as a dummy given that we fail to reject the null hypothesis of equal coefficient estimates for low and medium density MAZs (−0.080 for low density MAZs and −0.076 for medium density MAZs), both measured relative to the excluded high density MAZ category with a $\chi^2$ (1) value of 1.38 and because of the need to achieve analytic tractability for the MICE algorithm.

Excluding the number of persons living in the household, the estimated coefficients for all the variables are positive and significant at the 0.01 significance level. Thus, an increase in the value of any regressor increases the number of online purchases an individual makes in the past 30 days. However, the variable hhcount, with an estimated coefficient value of −0.04 and a p-value < 0.01 has the opposite effect, indicating that a unit increase in household size, with all other covariates constant, is associated with a decrease in the number of online purchases made.

Three of the most impactful covariates; all significant at the 0.001 significance level, are having at least a university degree; being in the working age population, defined as those aged 18 to 64, and belonging to a household that makes in excess of $100,000 income per year. Given that this is a NBRM, we exponentiate the coefficient estimates to ascertain the impact on online purchase given a unit change in the variable of interest. It is also worth pointing out that females purchase more online items compared to males, given the significant positive gender dummy. The vehicles explanatory variable, a count variable, has a positive, albeit marginal effect on the demand for online shopping. Having
a car is expected to be negatively correlated with online purchases given that meals, groceries or shopping mall trips could be made without the use of services such as Grubhub, Postmates, Instacart or Amazon. However, there is also a wealth effect that may be induced by car ownership, leading to increased demand for online purchases for at-home delivery. The net effect of these diametrically opposed factors is captured by vehicle’s coefficient estimate. We also report the estimate of the overdispersion parameter – both the transformed and the untransformed value. The estimate, relative to the 95% confidence interval confirms that the NBRM is more appropriate compared to the PRM.

We carry out a similar regression using Poisson regression model (PRM). Subtle differences are observed for the coefficient estimates across the two regression models though the coefficient estimates of each regressor for the Poisson regression relative to the negative binomial are of the same magnitude. In every instance, the associated p-values for the Poisson regression coefficient estimates are larger, relative to those from the NBRM. This is an indication of efficiency losses for the PRM. These estimates are subsequently used to create a fitted regression that generates the demand for online delivery services for each observation. Fig. 3 reveals that the assumption of no unobserved heterogeneity of the Poisson model may not be valid – a finding earlier corroborated by the summary data of the observed data where the variance is larger than the mean. As seen in Fig. 3, the Poisson model under-predicts the zeros and over-predicts the count data with magnitudes less than ten (Long and Freese, 2005).

4.1. Empirical Bayes estimates vs. imputed estimates

The observed demand data using data restricted to Florida, the NBRM overdispersion parameter and the predicted estimates are subsequently used to generate the Empirical Bayes (EB) estimates. The steps used for generating the EB estimates are outlined below:

1. Predicted demand: Using the NHTS survey, run a nationwide (excluding Florida) negative binomial regression for the deliveries at the individual level. Estimate the beta’s and the dispersion (phi).
2. Observed demand: Use the NHTS samples in Florida as the observed delivery demand at the individual level, match observed data to comparable synthetic individuals. We create classifiers based on the merged household and individual attributes with the observed demand determined by the mode of the data for each class.
3. Empirical Bayes: Calculate EB estimates using the formulas provided in Section 3.

Imputed estimates are obtained by running the MICE algorithm described in detail in Section 3. We then compare the EB estimates with the imputed demand for online goods by benchmarking both the EB and the imputed estimates relative to the ground truth – the NHTS Florida data on online delivery using two methods. One measure is fitting the predicted and imputed estimates of online purchases into distributions and then comparing them with that of the observed data. The second is generating population estimates of the weighted NHTS data and then comparing the figure with the cumulative predicted or imputed estimates obtained for synthetic individuals. Fig. 4 above shows the fitted distributions for both the imputed online delivery purchase estimates and the NHTS survey data. The EB estimate, not shown, performs worse compared to the imputed estimate given that the NB that provides the fit is characterized as NegBin (1, 0.1937). Relative to the imputed method, this provides a more pronounced contrast when both are compared with the observed NHTS data. As shown in Fig. 4, only a marginal difference is observed in the proportion of the imputed estimate relative to the observed data that fit in the 5 to 95% range of the NBRM distribution.

4.2. Predictive accuracy

With the volume of demand for online package deliveries for the Miami-Dade metropolitan statistical area (MSA) from the NHTS survey as the observed data, we measure the estimates relative to the observed data using the root mean square error (RMSE) and the mean absolute error (MAE). In addition, we also provide measures for the percentage of bad predictions which shows the number of instances for which the predicted demand falls outside a tolerance band and a scatter plot of imputed demand estimates, plotted against the observed data. Across the imputations, we obtain RMSE values ranging from a low of 15.73 to a high of 17.02, all lower figures relative to the 38.55 for the NBRM. To provide a different measure of how close the imputations are to each paired observed value, we calculate the percentage of bad predictions using a 30% tolerance value and obtain a bad prediction value of 14.2%, showing that one out of every seven predictions will be classified as bad. We do not report the percentage of bad predictions for the
NBRM because the figure is inflated by the conditional expectations used for the estimates particularly for households where no online purchases for at-home delivery were made. Beyond the use of a binary classifier for measuring the accuracy of the imputed estimates, we show a scatterplot in Fig. 5 that provides a panoramic assessment of the goodness of fit and insights on the errors for each paired data point.

In addition, using the person weights from the NHTS sample dataset, we generate population estimates of the demand for online delivery purchases for the Miami metropolitan area. We estimate that approximately 12.2 million online delivery purchases are made in a month in the Miami Metropolitan area. Using proportional population representation of the three-county area – Miami-Dade, Palm Beach and Broward Counties - in 2010, we estimate that Miami-Dade County accounts for about 5.56 million of these purchases. This figure is closer to the 5.93 million total imputed estimates obtained for Miami-Dade County compared to the 4.72 million obtained through the EB estimates.

We would like to point out that only one step of the updating process is carried out in generating the EB estimates. The NBRM demand estimates will always be positive even though 43% of individuals did not make any online purchases within the 30-day window referenced. And given that we are using conditional expectations, no variability will be observed in the predicted online purchases for at-home delivery as it is the case for the observed data for individuals with similar attributes.

5. Conclusion and areas for further research

The meteoric rise in urban goods delivery due to online shopping and app-based on-demand delivery services highlight the importance of this study, which at present is sorely lacking. Both the private sector and the public sector need accurate information on the demand for delivery services to create and coordinate more efficient, cost effective and societal beneficial systems for the delivery of goods and services to consumers. Strategies that implement mixed urban transport solutions, offered either by the private or the public sector such as integrating passengers and freight services are difficult to implement in the absence of on demand delivery data. In addition, the ability of cities to respond, in near real time, to situations such as those created by the COVID-19 pandemic may be diminished without the demand estimates – for example, better management of curbside spaces given the increase in the volume of online purchases. We provide a first step towards addressing these challenges by estimating the magnitude of online delivery purchases as well as variations across fine grained geographical areas. The research is particularly important because the more precisely goods and

Fig. 4. Fitted distribution for imputed estimates (top) relative to the observed NHTS survey data (below).
service providers can estimate demand, the better they can create competitive delivery systems that reduce costs and improve efficiency, especially in the last mile of the delivery process. Apart from the private sector, public sector officials could use the estimates in creating more robust delivery systems that explicitly tackle equity issues by implementing transportation solutions with both efficiency and equity objectives.

The current effort presents two methods to generate small area estimates of the demand for goods delivery using as a starting point the 2017 NHTS data coupled with synthetic household and individual records obtained from the South East Florida Regional Planning Model (SERPM). We first characterize the nature of the demand by fitting the NHTS observed data into distributions. Once the ideal distribution is identified, we subsequently use multiple approaches including an imputation method and an empirical Bayes method to generate estimated demand for MAZs in Miami-Dade County. The use of Miami-Dade County as a case study is informed by the availability of synthetic household and individual data at the MAZ level obtained from the SERPM. Of the two methods analyzed in this research, we find that the imputed approach provides better estimates of the demand for online delivery in geographic areas as small as 0.0015 mile². In addition, we observe appreciable variability in the estimated demand across the MAZs.

While the research provides an understanding of the demand for goods delivery in the absence of rich data sources, it could be enriched along multiple dimensions. For example, it should be possible to generate online delivery estimates nationwide for small areas at the US Census block groups (CBG) level by creating and integrating an iterative proportional fitting algorithm in our analysis. Further, we have not addressed the heterogeneity of online goods purchased. This is partially an artifact of the manner in which the NHTS survey questions are worded and the inability to access proprietary data such as Postmates data or credit card transactions. Employing private sector data could address the nature of the online delivery goods as well as the temporal dimension of the demand, as these transactions are time stamped. Extensions to the present study could also relax the assumption that demand for online delivery originates at the MAZ level and instead, provide more realism by assuming that door to door deliveries are made. This scenario obliges us to address the last mile delivery issue.

**Author statement**

Findings of the study were reviewed and the final version of the manuscript was approved by all authors.

**Acknowledgments**

This research was supported by Ford Motor Company – University of Michigan Alliance Program Grant. The authors thank Ford’s Robotics and Mobility Research department for supporting the research effort.

**References**

Allen, J., Cherrett, T., Piecyk, M., Piotrowska, M., 2019. The logistics of parcel delivery: Current operations and challenges facing the UK market. In: Browne, M. (Ed.), Urban Logistics: Management, Policy and Innovation in a Rapidly Changing Environment. Kogan Page, London, pp. 144–166.

Cameron, A.C., Trivedi, P.K., 2013. Regression Analysis of Count Data, 2nd edition. Cambridge University Press, Cambridge.

Cao, X., 2009. E-shopping, spatial attributes, and personal travel. Transp. Res. Record J. Transp. Res. Board 2135, 160–169. https://doi.org/10.3141/2135-19.

Cao, X., 2012. The relationships between E-shopping and store shopping in the shopping process of search goods. Transp. Res. A Policy Pract. 46 (7), 993–1002. https://doi.org/10.1016/j.tra.2012.04.007.

Cao, X., Douma, F., Cleveland, F., 2010. Influence of E-shopping on shopping travel. Transp. Res. Record J. Transp. Res. Board 2157, 147–154. https://doi.org/10.3141/2157-18.

Cardenas, I., Borbon-Galvez, Y., Verlinden, T., Van de Voorde, E., Vanelander, T., Devulf, W., 2017. City logistics, urban goods distribution and last mile delivery and collection. Compt. Regul. Netw. Ind. 18 (1–2), 22–43.

Carlin, B.P., Louis, T.A., 2000. Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall/CRC Press, Boca Raton, FL.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., 2014. Multivariate Data Analysis, 7th ed. Pearson Education Limited, Essex, England.

Joerss, M., Schröder, J., Neuhaus, F., Klink, C., Mann, F., 2016. Parcel Delivery: The Future of Last Mile. McKinsey & Co. https://www.supplychain247.com/paper/parcel_delivery_the_future_of_last_mile.

Lachapelle, U., Burke, M., Brotherton, A., Leung, A., 2018. Parcel locker systems in a car dominant city: location, characterisation and potential impacts on city planning and consumer travel access. J. Transp. Geogr. 71, 1–14. https://doi.org/10.1016/j.jtrangeo.2018.06.022.

Lee, K.J., Carlin, J.B., 2010. Multiple imputation for missing data: fully conditional specification versus multivariate Normal imputation. Am. J. Epidemiol. 171 (5), 624–632.

Long, J.S., Freese, J., 2005. Regression Models for Categorical Outcomes Using Stata, 2nd ed. Stata Press, College Station, TX.

Nguyen, D.H., de Leeuw, S., Dullaert, W., Foubert, B.P., 2019. What is the right delivery option for you? Consumer preferences for delivery attributes in online retailing. J.
Bus. Logist. 40 (4), 299–321.
Punel, A., Ermagun, A., Stathopoulos, A., 2018. Studying determinants of crowd-shipping use. Travel Behav. Soc. 12, 30–40.
Rahman, M.A., Islam, M.A., Esha, B.H., Sultana, N., Chakravorty, S., 2018. Consumer buying behavior towards online shopping: an empirical study on Dhaka city, Bangladesh. Cogent Business Manag. 5 (1), 1514940.
Rosa, A., Yunita, D., 2020, May. Consumer decision on online food delivery. In: 5th Sriwijaya Economics, Accounting, and Business Conference (SEABC 2019). Atlantis Press, pp. 418–422.
Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York.
Singh, S., Srivastava, S., 2018. Moderating effect of product type on online shopping behaviour and purchase intention: an Indian perspective. Cogent Arts Human. 5 (1), 1495043.
Sousa, R., Horta, C., Ribeiro, R., Rabinovich, E., 2020. How to serve online consumers in rural markets: evidence-based recommendations. Business Horizons 63 (3), 351–362.
Unnikrishnan, A., Figliozzi, M.A., 2020. A Study of the Impact of COVID-19 on Home Delivery Purchases and Expenditures. Portland State University Working Papers Series.
vан Buuren, S., 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med. Res. 16 (3), 219–242.
Van Droogenbroeck, E., Van Hove, L., 2017. Adoption of online grocery shopping: personal or household characteristics? J. Internet Commer. 16 (3), 255–286.
Visser, J., Nemoto, T., Browne, M., 2014. Home delivery and the impacts on urban freight transport: a review. Procedia Soc. Behav. Sci. 125, 15–27. https://doi.org/10.1016/j.sbspro.2014.01.1452.
Wang, X.C., Zhou, Y., 2015. Deliveries to residential units: a rising form of freight transportation in the US. Transp. Res. Part C Emerg. Technol. 58, 46–55.
White, I.R., Royston, P., Wood, A.M., 2011. Multiple imputation using chained equations: issues and guidance for practice. Stat. Med. 30 (4), 377–399.
Yeo, V.C.S., Goh, S.K., Rezaei, S., 2017. Consumer experiences, attitude and behavioral intention toward online food delivery (OFD) services. J. Retail. Consum. Serv. 35, 150–162.
Zhou, Y., Wang, X., 2014. Explore the relationship between online shopping and shopping trips: an analysis with the 2009 NHTS data. Transp. Res. A Policy Pract. 70, 1–9. https://doi.org/10.1016/j.tra.2014.09.014.