NEW YORK METROPOLITAN TRANSPORTATION COUNCIL

FINAL REPORT FOR THE SEPTEMBER 11TH MEMORIAL PROGRAM

ANALYSIS OF SOCIOECONOMIC AND DEMOGRAPHICS METHODOLOGIES AND FORECASTING APPROACHES

JULIA AMARAL

September 2022

Table of Contents

1. Introduction

Reasonable and defensible forecasts are a critical part of NYMTC's regional transportation planning process. NYMTC has developed an Excel based socioeconomic and demographic (SED) forecasting system to forecast county-level employment and population through 2050. NYMTC has also developed an Excel based Zonal Allocation Process (ZAP) to allocate those forecasts to the Transportation Analysis Zone (TAZ) level for transportation modelling purposes.

The objective of this research is to evaluate new methods to update NYMTC's socioeconomic and demographic forecasts and recommend enhancement or alternatives to the existing 2050 Socioeconomic and Demographic (SED) forecasting model(s) for the 31-county New York Metropolitan Region that would enable NYMTC's staff to independently develop future SED forecasts and zonal allocations. The recommended methodology must be linked to NYMTC's New York Best Practice Model (NYBPM) and other regional models, and it must consider the influence of the current economic environment on future conditions, and structural changes in the economy that may be underway and ensure that regional land use trends and policies of NYMTC's members are reflected in the population, employment, households, and labor force forecasts.

The approach used in this research included a literature review of current practices and state-of-the-art forecasting methods. Two main components of the research are a literature review of papers in academic journals and a review of methodologies adopted by other MPOs in the United States which was obtained by reports published by MPOs and an e-mail survey. Figure 1 presents a schematic of the approach used in this report including the questions it aims to address to build to the evaluation and recommendation of methods.

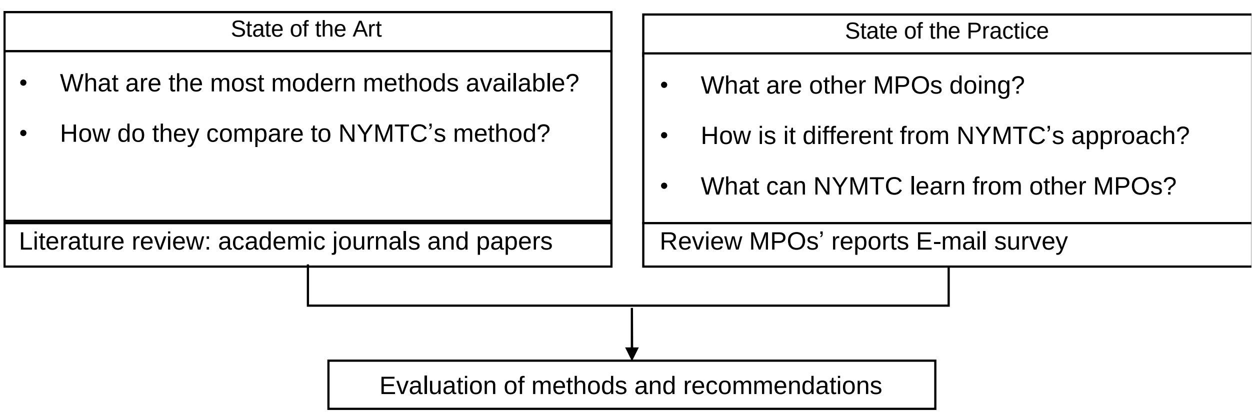


Figure 1: Approach of the study

This report summarizes the findings of both approaches in two sections, one dedicated for the literature review and the other dedicated for the review of the MPOs. Lastly, this report also includes a chapter with the recommendations for NYMTC based on lessons learned from the research.

2. Literature Review

Forecasting trends in population size, age structure, employment, regional distribution, and other socio-demographic variables are necessary in a variety of planning situations. For that reason, the literature is quite extensive and covers a wide range of methods, from large scale regions and national forecasting (Booth, 2006; Raftery et al., 2012; Mazzuco & Keilman, 2020) to small area population forecasts that are extremely useful in the context of urban planning. The terminology "small area" forecasting refers to the level of aggregation of the forecasting. While large scale forecasting focuses on forecasting total numbers for a single demographic area (e.g, country, or state), small area forecasting aims for more disaggregate results, in which sociodemographic variables are known for individual smaller areas that compose a region. Hence, small area forecasting techniques are widely used for planning purposes at metropolitan level. Due to the purpose of NYMTC's forecasting efforts, the focus of this literature review will be on small area forecasting.

Since forecasting is useful for a variety of planning purposes, the literature offers a variety of studies on forecasting socio-demographic variables without a specific focus on the purpose of obtaining such information. In those cases, the objective of the research is the forecasting itself. Additionally, there is also research focused specifically on transportation planning in which forecasting sociodemographic comes as a crucial step of the process. Traditional transportation models (i.e, four-step model, agent-based model) use forecasts as an input for modeling, whereas the called integrated transportation-land use models explore the interrelation between land use and transportation infrastructure to generate socioeconomic forecasts. For this reason, this literature review covers studies of core forecasting techniques and integrated transportation land use models.

Wilson et al. (2021) produced a review of the current small area forecasting methods and classified them into eight categories: 1) extrapolative and comparative methods, 2) simple cohort-component methods, 3) model averaging and combining, 4) econometric models, 5) housing-led population projections, 6) 'downscaling' and disaggregation approaches, 7) small area microsimulation, and 8) machine learning methods.

This literature review will use the categories defined by Wilson et al. (2021) and will add a nineth category to discuss integrated transportation land use models.

2.1 Extrapolative and Comparative Methods

These simple methods are usually employed to estimate population totals instead of specific groups (e.g., by age or by sex). The application is very straightforward and consists in identifying growth trends (usually linear or exponential). Although based solely on past trends, their advantages include minimal data requirements, simple calculation, and accuracy that is frequently comparable to, if not better than, more detailed and complex methods. However, the accuracy of results is dependable of a certain set of conditions. It cannot incorporate the influence of external factors that could affect the trends observed in the past such as economical changes or pandemics. For this reason, they must be used with caution and are more appropriate for short term forecasting.

For decades, practitioners and researchers are applying extrapolative and comparative methods in their forecasts (White, 1954; Isserman, 1977; Smith, 1987). Their popularity decreased in the most recent years with the advent of more powerful computers that allowed for the implementation of more complex methods of forecasting. However, recent studies compare simple extrapolative methods to other methods and show that their accuracy is comparable.

Rayer (2008) evaluated simple methods for forecasting the total populations of counties in the US. Among models tested, linear extrapolation was found the most accurate by a small margin. Wilson (2015) compared the accuracy of several extrapolative and comparative methods to estimate small area population forecasts in Australia, New Zealand, England, and Wales. The best performance was given by a constant share of growth model, and a constant share-of-population model.

2.2 Cohort-Component Methods

The simple cohort-component method is based the assumption that age-specific vital rates and migration rates of the recent past will continue unchanged into the near future (Hamilton & Perry, 1962). Also known as the Hamilton-Perry model, it has been implemented extensively for small area forecasting and has some variations.

For example, Swanson et al. (2010) adjusted cohort change ratios to match population estimates early in the forecast horizon and applied lower and upper bound limits to avoid unwanted large population growths or reductions in census tracts and block groups. Another example is Baker et al. (2014) that also used it to forecast urban census tracts population but combined it with spatial weighting applied to preliminary forecasts.

Wilson (2016) evaluates alternative cohort component models for small area forecast and compares their performances when applied isolated or when total population is constrained by an extrapolative method. Five versions of the method are considered:

- bi-regional cohort-component model
- cohort-component model using net migration numbers
- cohort-component model using net migration rates
- composite net migration cohort-component (mix of net migration numbers and net migration rates)
- Hamilton-Perry shortcut cohort model

Ten retrospective forecasts are produced (five methods constrained + five methods unconstrained) and compared with population estimates for local areas in New South Wales, Australia. Results indicate that constraint versions perform significantly better, the constrained bi-regional cohort component method gave the lowest errors.

2.3 Model Averaging and Combining

Averaging and combining are often found to reduce errors in forecasts. Even though the benefits of model averaging and combining are highlighted in the literature for decades (Bates & Granger, 1969; Clemen, 1989), such approaches are not usually applied. Goodwin (2009) discusses that combining models uses more data than any singular model, making errors from different models offset each other to some degree.

The literature provides examples where averaging produced better results than single methods. Rayer (2008) created forecasts for US counties using five simple methods and found that a mean of all five methods produced smaller errors than best individual methods. Similarly, Rayer and Smith (2010) found that averaged forecasts performed well to forecast population in sub-county areas in Florida.

Combining methods also shows good results in the literature. As an example, the before mentioned results from Wilson (2016) that cohort-component methods with total population constrained by extrapolative methods provided better results than cohort-component methods individually.

2.4 Econometric Models

Researchers discuss the inclusion of socioeconomic variables and spatial relationships in small area forecasting because demographic patterns are not isolated. Demographic trends are influenced by migration and short-term mobility, as well by characteristics that are often similar in nearby areas—social norms, type of housing, regulations, culture, politics, etc. (Wilson, 2021)

To capture the effect of more characteristics that influence forecasting, models include explanatory variables related to demographics, socioeconomic characteristics, transportation accessibility, amenities, and land development. However, studies show that regression models for forecasting total populations from a set of independent variables generally do not produce substantially more accurate forecasts than simple extrapolative models. For example, geographically weighted regression model applied to minor civil divisions in Wisconsin performed slightly less accurate than several simple extrapolative forecasts (Chi & Wang, 2017).

2.5 Housing-Led Population Projections

In this approach, population is calculated as the number of occupied residential units multiplied by the average household size. An advantage of this method is that, due to policies and regulations in urban areas, the existing and the future number of private dwellings which will be built in the future (approx. 10 years) is known with a reasonable level of assurance. It is a simple method that provides good accuracy to forecast total population in small areas. Its main disadvantage, however, is that it does not generate detailed forecasts for age groups or sex, for example (Wilson, 2021).

The housing-unit model is regularly used by private sector demographers, the public sector, and academia, often in combination with other models. Hauer et al. (2015), for example, used a Linear/Exponential model to forecast dwelling numbers, and then estimated population in sub-county areas in Georgia. Another alternative is to use household-led forecasts combined with a cohort-component model, allowing changes in population age structure to influence average household size. In this approach applied by Simpson (2017) for local areas in the United Kingdom, instead of assuming an average household size, it is one of the model's outputs.

2.6 'Downscaling' and Disaggregation Approaches

These methods are also called "top-down" forecasting, when small area forecasts are produced based of larger areas' forecasts. For NYMTC, these methods are relevant for the zonal allocation step. In the literature, a common application is to downscale population forecasts to small cell grids. For example, Breidenbach et al. (2019) generated age–sex population forecasts for 1 km^2 grid squares in Germany between 2015–2050.

The main advantage of disaggregation is that population data on a regional level is usually readily available—national census, for example. However, Wardrop et al (2018) discuss challenge of finding reliable input data, and how it affects the accuracy of the final forecasts for smaller areas.

The most common technique uses disaggregation weights. The large area *M* (representing variable, e.g., population) are distributed among the small areas *i* using equation (1), where w_i are the disaggregation weights of the small areas.

$$m_i = M \frac{w_i}{\sum w_i} \quad (1)$$

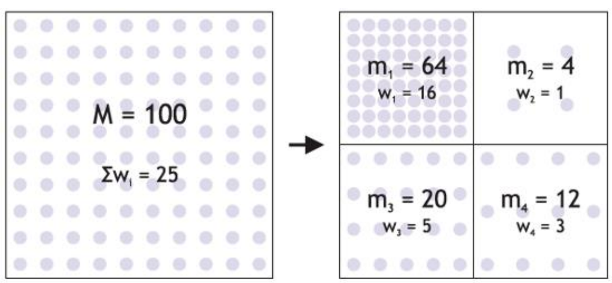


Figure 2: Weighted spatial disaggregation schematics

The simplest method is to use surface area as weights, but other variables could be incorporated to improve the accuracy of the disaggregation, such as street length, number of dwellings, or built area. Simbera 2020 uses open-source machine learning tools to compute weights for weighted population disaggregation. The results show that land use and buildings data were the most influential variables for population disaggregation.

2.7 Small Area Microsimulation

Microsimulations model at the scale of individuals by definition. For that reason, they require considerable amounts of inputs and data preparation to provide rich details across various population characteristics (Wilson, 2021). The data required depends on the simulation approach and the variables modeled, but basically must include information of the current population at an individual level (e.g., household location, income, employment, educational level), and any additional information that could help model natality and mortality rates (e.g., age, life expectancy, access to health care) and migration (e.g., land cost, housing market, economic developments). The output of a microsimulation contains information for all the variables modeled for each individual of the population.

Examples are SMILE (Simulation Model for the Irish Local Economy), a simulation developed to forecast population for small areas of Ireland for the periods between 1991 and 2002, (Ballas et al., 2005), and the simulation developed by Marois and Bélanger (2014) to forecast the populations of municipalities within the metropolitan region of Montreal, Canada from 2006 to 2031. The forecast accuracy is reasonable; mean absolute error obtained for both simulations were 6.4% and 3.4% respectively.

2.8 Machine Learning

Machine learning methods are popular and successful in a variety of applications. However, in forecasting these methods seem to perform worse than traditional statistical methods. Makridakis et al. (2018) evaluated and compared the performance of forecasts produced by machine learning methods (Long Short-Term Memory models, Bayesian Neural Networks, and Regression Trees) against traditional statistical methods. Results show that traditional methods produced better accuracy even though they had lower computational requirements than machine learning methods.

In essence, even though machine learn methods for forecasting are the state-of-the-art research, their poor performance in comparison with traditional forecasting methods combined with the higher computational requirements does not make them advantageous for real-life applications.

2.9 Integrated Transportation Land Use Model

The relationship between urban development and transportation is complex and linked to each other; they are also inextricably linked to other urban processes such as macroeconomic development, interregional migration, demography, household formation, and technological innovation. Integrated models of urban land use and transportation are decision support tools for urban planning. They simulate the two-way interaction between land use and transportation to estimate the potential implications of land use and transportation policies.

These models incorporate land use and transportation infrastructure, and often environmental effects, into the planning and management of metropolitan areas. This type of models emerges from the growing concern that building, for example, new highways will increase travel demand, emissions, and will induce land development, making it virtually impossible to "build our way out" of congestion. This broad approach is reshaping the planning and policy context for metropolitan regions. The Federal Highway Administration, Federal Transit Administration, and the Environmental Protection Agency formed the Travel Model Improvement Project (TMIP) in response to the growing concern on the limitations of traditional transportation models.

According to Weatherby (1995) the main guidelines from the TMIP are:

- Prioritize random utility-based models,
- Use a behavioral basis describing the principal actors involved in urban development and transportation,
- Give emphasis to models for policy analysis, planning, and sensitivity testing,
- Recognize the varying temporal and geographic scales relevant to urban development,
- Move towards more disaggregate data,
- Use an interdisciplinary approach,
- Develop modular models,
- Increase the use of GIS,
- Test the effects of transportation on land use.

In 1999, Wegener and Fürst introduced the concept of the land use transport feedback cycle (Figure 3), recognizing that mobility and location decisions are intertwined and, as a result, transportation and land-use planning must be coordinated. It basically describes that the distribution of land uses, i.e., residential, industrial, commercial, define the location of households and activities such as working and leisure, and the distribution of activities require trips in the transportation system, then travelers make travel decisions based on car and transit availability, cost, travel time, etc. Finally, the travel decisions define the attractiveness of locations and result in changes in buildings such as new investments, upgrades, or even demolitions, and these changes end up affecting the distribution of land uses in the city.

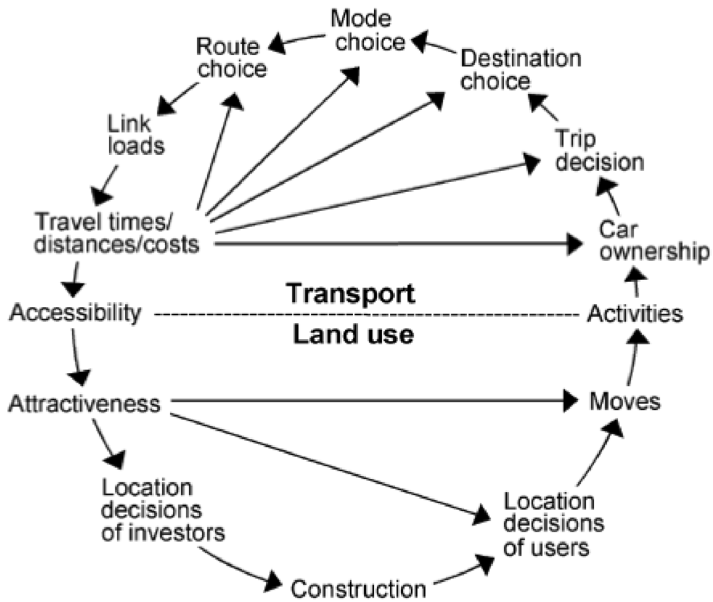


Figure 3: The land use transport feedback cycle (Wegener and Fürst, 1999).

In general, there are three types of approaches for predicting effects of land-use transportation policies. The first is known as 'expressed preference'; it consists of asking people how they would react to changes such as higher transportation costs or land use restrictions. The second option is called 'revealed preference’ and consists of drawing conclusions from observed human behavior. The third category of approaches includes mathematical models that simulate the human decision-making process and its effects. Although all three options have drawbacks, mathematical models are the only way to foresee unknown circumstances and identify the effect of a single component while holding all other factors constant (Wegener and Fürst, 1999).

Wegener (2021) proposes a classification of models according to the way they implement the feedback from transportation to land use: (i) spatial interaction location models, and (ii) accessibility-based location models. Spatial interaction location models forecast the locations of production and consumption in the metropolitan area using a multi-industry, multiregional input-output framework in which households are represented as industries producing labor and consuming commodities. The equilibrium between transportation costs and land and commodity prices are achieved by iterating between the land-use and transportation components of the models. PECAS (Hunt and Abraham, 2005) is a current example of a spatial interaction location model.

The second category of land-use transport models, accessibility-based location models, use accessibility measures to predicts possible spatial interactions. Accessibility measures can range from simple indicators, such as the distance to the nearest bus station or highway exit, to complex indicators that measure the easiness to reach various points of interest. These models use discrete choice models with multi-attribute utility functions to predict household and firm locations; accessibility indicators are combined with other attributes of potential locations to indicate their attractiveness from the perspective of people searching for a residential location or companies searching for a business location. A recent example of an accessibility-based location model is UrbanSim (Waddell 2002).

According to the survey conducted with select MPOs, UrbanSim is the most popular integrated land use transportation tool used by MPOs in the United States. Although it was originally created as an opensource tool at the University of Berkeley, over the years it has consolidated itself as a commercial tool, firstly marketed by Autodesk and currently marketed by UrbanSim Inc.

The UrbanSim framework contains many components. There are three main focuses: employment, household, and real estate. The model predicts the evolution of these entities using annual steps to predict the movement and location choices of businesses and households. The land use model is interfaced with a travel model system to deal with the interactions of land use and transportation, in which accessibility to opportunities (employment, shopping, etc.) is measured by the travel impedance to reach these opportunities through all possible modes. The outputs are thorough and include tables of individual households and persons, jobs, parcels, buildings, with their attributes updated each simulation year if they have been modified by the model system (Waddell et al., 2018)

In sum, transportation land use models are preferred by many agencies for the extensive number of outputs produced and the capacity of predicting land use/transportation effects of policies, a more holistic approach than modeling each separately. However, these models require large amounts of disaggregate data that could make their application unfeasible not just because of costs of implementation but also because of unavailability of key inputs. With the rapid digitalization of processes, the expected trend is that such data becomes more easily accessible and perhaps the effort and cost of building integrated transportation land use models will reduce significantly, making them more appealing.

3. Review of MPOs' Methodologies

3.1 Introduction

The objective of this task is to provide a comprehensive analysis of peer MPOs' practices for socioeconomic and demographic forecasting (SED) and zonal allocation process (ZAP). The intent is to deliver an overview of the current trends and identify potential methods that could be adopted by NYMTC. MPOs were selected to participate based on the size of the metropolitan region of their jurisdiction. The analysis was accomplished by reviewing the publicly available material that MPOs publish online and by a survey that was conducted by email. Ten MPOs responded to the survey:

- Atlanta Regional Commission (ARC)
- Association of Bay Area Governments – Metropolitan Transportation Commission (ABAG-MTC)
- Baltimore Regional Transportation Board (BMC)
- Chicago Metropolitan Agency for Planning (CMAP)
- Delaware Valley Regional Planning Commission (DVRPC)
- Houston-Galveston Area Council (H-GAC)
- Metropolitan Washington COG (MWCOG)
- North Central Texas COG (NCTCOG)
- North Jersey Transportation Planning Authority (NJTPA)
- Southern California Association of Governments (SCAG)

3.2 The Survey

The questions of the survey targeted the SED and ZAP practices at the peer MPOs, including frequency, tools used, and necessary resources. Six questions were asked:

1) How often do you update the base-year and horizon year SED forecasts?
2) What methodology and software do you use for your regional / county level long term SED forecasts?
3) What software or methodology is used for the zonal allocation?
4) How many milestone years are included in your SED forecasts?
5) If the forecast is done by in-house staff, how many staff members are involved in the process (full-time and part-time)?
6) If the base-year forecasts are developed by consultant, are you able to update the forecast with in-house staff for intermediate years?

3.3 Results

Table 1 summarizes the population of the peer MPOs in the baseline and horizon years and the frequency of forecast updates produced by each MPO. The most popular update frequency is 4 years, and except for Baltimore that produces forecasts as needed, all MPOs produce forecasts with a regular frequency. Another popular response among the surveyed MPOs is to produce a first cycle of forecast every new full decennial census release and produce smaller updates due to local conditions on a more frequent basis.

Table 1: Peer MPOs' population and frequency of updates

MPO	Baseline year	Baseline Population	Horizon year	Horizon Population	How often forecasts are updated
New York Metropolitan Transportation Council	2017	22.7 Million	2055	25.9 Million	every 4 years
South California Association of Governments	2012	22.6 Million	2040	26.7 Million	every 4 years
Association of Bay Area Governments	2015	7.6 Million	2050	10.5 Million	every 4 years
Chicago Metropolitan Agency for Planning	2015	7.3 Million	2050	10.6 Million	every 8 years
North Jersey Transportation Planning Authority	2017	6.7 Million	2050	7.7 Million	every 4 years
Houston-Galvestone Area Council	2015	6.5 Million	2045	10.5 Million	every 2 years
Atlanta Regional Commission	2015	5.7 Million	2050	8.6 Million	every 3 to 4 years
North Central Texas Council of Governments	2017	5.7 Million	2045	11.2 Million	every 4 to 5 years
Delaware Valley Regional Planning Commission	2015	5.6 Million	2050	6.4 Million	every 4 years
Metropolitan Washington Council of Governments	2015	5.5 Million	2045	6.8 Million	every 5 years
Baltimore Regional Transportation Board	2015	2.8 Million	2050	3.1 Million	as needed

Table 2 summarizes whether the MPOs use the services of a consultant, whether they use a land use model to produce their forecasts, and the allocated staff for SED forecasting and ZAP. For the sake of simplicity, the table mark as "consultant" all MPOs that work with consultants to produce the forecasts, even if they just use tools provided by consultants and the forecasts are produced by in-house staff. Most of the MPOs produce updates of their forecasts in-house even if they hire a consultant for an initial forecast. The survey also showed that land use models are still not broadly used; only four MPOs reported to use land use models as part of their modeling efforts. Cohort-component models are still the most popular method to produce forecasts.

Table 2: Peer MPOs' staff needs, and consultant and land use adoption

MPO	Forecasts produced in house or by consultant	Updates produced in-house or by consultant	Land Use Model	Number of staff members (FTE)
New York Metropolitan Transportation Council	Consultant	Consultant	No	2
South California Association of Governments	In-house	In-house	No	4-10
Association of Bay Area Governments	Consultant	In-house	Yes	5
Chicago Metropolitan Agency for Planning	Consultant	In-house	Yes	2.5
North Jersey Transportation Planning Authority	Consultant	Consultant	No	1
Houston-Galvestone Area Council	In-house	In-house	Yes	5
Atlanta Regional Commission	Consultant	In-house	No	4
North Central Texas Council of Governments	Consultant	In-house	No	4
Delaware Valley Regional Planning Commission	Consultant	In-house	Yes	1
Metropolitan Washington Council of Governments	In-house	In-house	No	5
Baltimore Regional Transportation Board	Produced by smaller agencies	Produced by smaller agencies	No	-

More details about the MPOs can be found in one-pager summaries included in this document. Each summary includes an overview of the methods used by the MPO, key information regarding the jurisdiction of the MPO, tools used in the forecast, variables produced, links to reports, etc. In addition, the information is also consolidated in a excel table, where the reader can find the main information obtained about the MPOs.

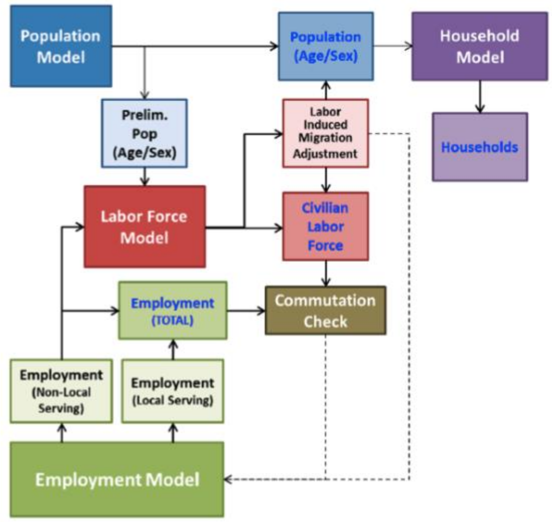
The information on the one-page summaries were collected though the e-mail survey and MPOs' reports. Table 3 lists all the fields of the one-pagers, the most probable source (survey, reports, or combination), and any additional observations referring to that field.

Table 3: Fields included in the one-pager summaries.

Name of the Field	Source	Obs.
Overview	Combination	This field is a brief explanation of the MPOs approach
Counties under MPO's Governance	Reports	
Baseline Year	Reports	
Population at Baseline Year	Reports	
Horizon Years	Combination	
Population at Horizon Year	Reports	
Forecast Years	Reports	
Consultant	Combination	Refers if the MPO used a consultant in any step of the forecasting process
Updates produced by	Survey	
Frequency of Updates	Survey	
Date of Most Recent Report	Reports	Date of the most recent report available online at the time of the research
Socioeconomic and Demographic (SED) Forecasting Tool	Survey	
Zonal Allocation Process (ZAP) Tool	Survey	
Land Use Model	Combination	Indicates if the MPO uses any type of integrated land use model for the transportation model
Staff	Survey	Number of staff in the forecasting efforts
Number of Variables Produced	Report	The number of variables produced was gathered from reports. In the case where the variables produced were not explicitly mentioned, they were estimated based on the description of the outputs and the forecasts reported
Point of Contact	Survey	Name of the respondent of the survey
Email	Survey	
Links	NA	Links to the reports and general websites with forecasting information

New York Metropolitan Transportation Council (NYMTC)

Overview: Although there are ten counties under NYMTC's governance, the forecasts are produced for 31 counties through a Forecasting Working Group that includes representants of multiple agencies in the area. The forecasting process is composed by two main steps; firstly, population, household and employment variables are estimated for counties, and secondly, they are used as control totals for the allocation into TAZs. There are three models (employment, population, TAZ allocation) based on Excel. The figure shows the relation between the employment and population models. Sixteen variables are produced at TAZ level as a result of the forecasting effort, and they are used as an input for the New York Best Practice Model that does the transportation modeling of the region.



Counties under MPO's Governance: New York, Kings, Queens, Bronx, Richmond, Nassau, Westchester, Rockland, Suffolk, Putnam

Baseline Year: 2017	**Population at Baseline Year**: 22.7 million
Horizon Year: 2055	**Population at Horizon Year**: 25.9 million

Forecast Years: 2020, 2025, 2030, 2035, 2040, 2045, 2050, 2055

Consultant: WSP and Urbanomics

Updates produced by: Consultant	**Frequency of Updates:** every 4 years

Date of Most Recent Report: October 2020

Socioeconomic and Demographic (SED) Forecasting Tool: Excel

Zonal Allocation Process (ZAP) Tool: Excel

Land Use Model: No

Staff: 2

Number of Variables Produced: 16

- Population (Total, Household, Institutionalized Group Quarter, Homeless, Other Group Quarters)
- Households (Number of Households, Avg. Size, Avg. Income)
- Employment (Total, Retail, Office, Earnings per worker)
- University and K-12 School Enrollment

Point of Contact: Larisa Morozovskaya	**Email:** larisa.morozovskaya@dot.ny.gov

Links:

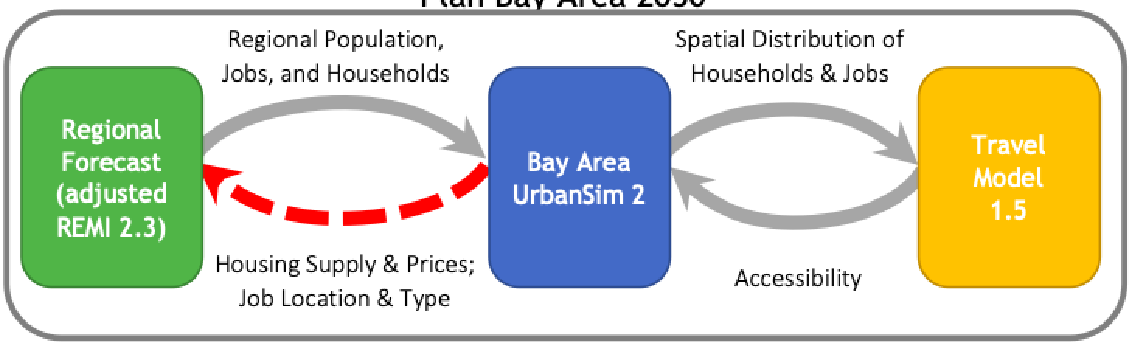
- https://www.nymtc.org/DATA-AND-MODELING/Socioeconomic-and-Demographic-SED-Forecasts/2055-Forecasts
- https://www.nymtc.org/Data-and-Modeling/SED-Forecasts/2050-Forecasts
- https://www.nymtc.org/Portals/0/Pdf/SED/2055%20SED/Technical%20Memo%202_Model%20Update%20to%202017%20Baseline.pdf?ver=upAxDCc8BgPAVryEdUtjgw%3d%3d
- https://www.nymtc.org/Portals/0/Pdf/SED/2050%20SED/160107-T6-ZAPMethod-Final.pdf?ver=2016-01-22-130437-170

Southern California Association of Governments (SCAG)	
Overview: SCAG uses a cohort component population projection model built in Excel for county and region-level population forecasts. Households are derived from the population projection using headship rate scenarios. Separately, a shift-share model is used to project regional employment – this model is also custom built and mostly conducted in Excel. The following step is zonal allocation; county level growth is disaggregated to jurisdictional and traffic analysis zonal levels by using general plans, specific plan, and other data sources. The current model was developed in SAS, but it is transitioning to other languages such as Python.	
Counties under MPO's Governance: Imperial, Los Angeles, Orange, Riverside, San Bernardino, Ventura	
Baseline Year: 2012	**Population at Baseline Year**: 22.6 million
Horizon Year: 2040	**Population at Horizon Year**: 26.7 million
Forecast Years: 2020, 2025, 2030, 2035, 2040	
Consultant: -	
Updates produced by: In-house	**Frequency of Updates:** every 4 years
Date of Most Recent Report: September 2020	
Socioeconomic and Demographic (SED) Forecasting Tool: proprietary models built in excel	
Zonal Allocation Process (ZAP) Tool: proprietary models built in SAS and Python	
Land Use Model: No	
Staff: 4-10 full-time	
Number of Variables Produced: more than 50 • Household (type, size, income, housing tenure) • Employment (2-digit NAICS and occupation) • population (occupation, age, sex, ethnicity, educational level)	
Point of Contact: Ying Zhou	**Email:** zhou@scag.ca.gov
Links: • https://scag.ca.gov/data-tools-forecasting • https://scag.ca.gov/regional-forecasting • https://scag.ca.gov/subarea-forecasting • https://scag.ca.gov/sites/main/files/file-attachments/0903fconnectsocal_demographics-and-growth-forecast.pdf?1606001579	

Association of Bay Area Governments (ABAG)

Overview: The forecasting procedure involves two main steps; (i) regional forecast growth and (ii) land use model. For the regional growth forecast, ABAG uses the Regional Economic Models, Inc. (REMI) Policy Insight+ (or PI+) tool, which accounts for industry structure and competitiveness relative to other regions, propensity to export, and population and labor market structure, to forecast the growth in jobs by industry, housing units and population in the Bay Area. County level forecasts are obtained by aggregation of UrbanSim 2' results, that forecast jobs and households at TAZ level.

Plan Bay Area 2050



Counties under MPO's Governance: Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, Sonoma

Baseline Year: 2015	**Population at Baseline Year**: 7.6 million
Horizon Year: 2050	**Population at Horizon Year**: 10.5 million

Forecast Years: 2020, 2025, 2030, 2035, 2040, 2045, 2050

Frequency of Updates: Previous forecast was produced in 2018

Consultant: Regional Economic Models, Inc. (REMI)

Updates produced by: In-house	**Frequency of Updates:** every 4 years

Date of Most Recent Report: October 2021

Socioeconomic and Demographic (SED) Forecasting Tool: REMI Policy Insight + 2.3.1

Zonal Allocation Process (ZAP) Tool: UrbanSim 2

Land Use Model: UrbanSim 2

Staff: two full-time staff were involved in developing the Regional Forecast, and 2.5 full-time staff were involved in developing the small-area allocation

Number of Variables Produced: 34

- Employment by 11 sectors
- Population by 15 age groups and 4 ethnic characteristics
- Households by 4 income levels

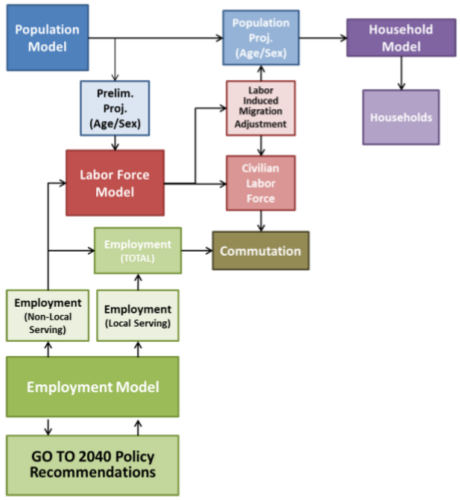
Point of Contact: Bobby Lu	**Email:** blu@bayareametro.gov

Links:

- https://www.planbayarea.org/sites/default/files/documents/PBA50_Forecasting_and_Modeling_Report_Oct2021.pdf
- https://abag.ca.gov/our-work/land-use/plan-bay-area-2050
- https://abag.ca.gov/our-work/land-use/forecasts-projections

Chicago Metropolitan Agency for Planning (CMAP)

FIGURE 2-1. FLOW CHART FOR FORECASTING APPROACH



Overview: The forecasts were previously produced by Louis Berger (consultant). The MPO is on the first cycle of a new approach that builds the demographic model on the previous tool produced by Louis Berger. The documentation provided on their website refers to the previous models. CMAP also relies on advisory from a demographer from the University of Wisconsin for the development of this new tool. Together they are producing an in-house cohort component model.

In addition, they are also working with consultant EBP for their economic model. The employment model uses an averaging process that combines independently developed forecasts from Moody's, and the U.S. Bureau of Labor Statistics (BLS).

Documentation on the previous model, states that household forecasts are derived by applying age/sex-specific household formation (headship) rates to the age/sex profiles obtained from the population forecasts.

Counties under MPO's Governance: Cook, DuPage, Kane, Kendall, Lake, McHenry, Will

Baseline Year: 2015	**Population at Baseline Year**: 7.3 million
Horizon Year: 2050	**Population at Horizon Year**: 10.6 million

Forecast Years: 2020, 2025, 2030, 2035, 2040, 2045, 2050

Consultant: University of Wisconsin/EBP

Updates produced by: In-house	**Frequency of Updates:** every 8 years

Date of Most Recent Report: November 2016

Socioeconomic and Demographic (SED) Forecasting Tool: Demographic is an in-house cohort component model based on work previously done by a consultant (Louis Berger). Economic is a modified version of Moody's produced by consultant EBP. Both implemented in R.

Zonal Allocation Process (ZAP) Tool: UrbanSim

Land Use Model: UrbanSim

Staff: 2.5 FTE over the past two years

Number of Variables Produced: more than 50

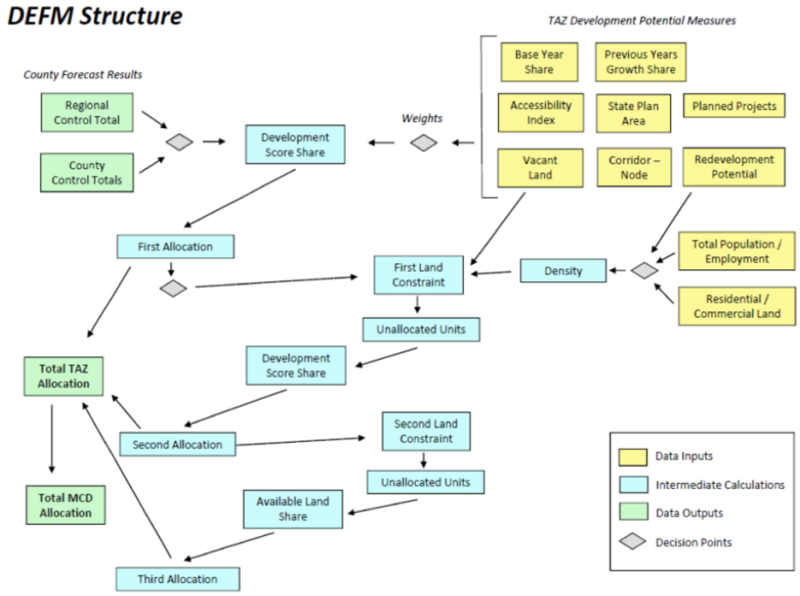
- Population by 18 age groups
- Population by race and ethnicity (white, Hispanic, black, Asian, other)
- Employment by 20 industry sectors (2-digit NAICS)
- Households (By number of persons, By 3 groups of age of householder, By number of workers, By sex, by prescribed age ranges, By 11 income ranges)
- Non-institutionalized group quarters population (By group quarters population, by sex, by prescribed age ranges)

Point of Contact: David C. Clark	**Email:** dcclark@cmap.illinois.gov

Links:

- https://www.cmap.illinois.gov/data/demographics/population-forecast
- https://datahub.cmap.illinois.gov/dataset/89f66569-5f51-4c14-8b02-5ecc1ca00909/resource/a812de2f-d465-47f2-87df-0427e81da2cf/download/CMAPSocioeconomicForecastFinal-Report04Nov2016.pdf

North Jersey Transportation Planning Authority (NJTPA)



Overview: NJTPA works with NYMTC to determine the county and regional forecasts. The zonal allocation process uses an in-house Excel based model to allocate the county control total forecasts to the local TAZ level. The model uses information such as current land use, zoning, accessibility, historical growth, and known project developments to perform the allocation. The figure shows the general schematics of NJTPA's Demographic and Employment Forecast Model (DEFM).

Counties under MPO's Governance: Bergen, Essex, Hudson, Hunterdon, Middlesex, Monmouth, Morris, Ocean, Passaic, Somerset, Sussex, Union, Warren

Baseline Year: 2017	**Population at Baseline Year**: 6.7 million
Horizon Year: 2050	**Population at Horizon Year**: 7.7 million

Forecast Years: 2020, 2025, 2030, 2035, 2040, 2045, 2050

Consultant: forecasts produced with NYMTC

Updates produced by: Consultant	**Frequency of Updates:** every 4 years

Date of Most Recent Report: September 2021

Socioeconomic and Demographic (SED) Forecasting Tool: NYMTC's models

Zonal Allocation Process (ZAP) Tool: proprietary model developed in excel

Land Use Model: No

Staff: 1 full-time

Number of Variables Produced: 4

- Population
- Employment
- Households
- Household Size

Point of Contact: Bob Diogo	**Email:** bob@njtpa.org

Links:

- https://www.njtpa.org/Data-Maps/Modeling-Surveys/Population-Jobs-Model.aspx
- https://www.njtpa.org/Data-Maps/Modeling-Surveys/Land-Use-Model.aspx
- https://www.njtpa.org/Data-Maps/Demographics-GIS/Forecasts.aspx
- https://www.njtpa.org/NJTPA/media/Documents/Data-Maps/Demographics-GIS/Forecasts/DEFM_User_Guide_June2011revision.pdf
- https://www.njtpa.org/NJTPA/media/Documents/Data-Maps/Demographics-GIS/Forecasts/E-2050-Demographic-Forecasts.pdf

Houston-Galvestone Area Council (H-GAC)	
Overview: H-GAC uses an integrated transportation land use model, like UrbanSIM. The forecast is produced in phases: 1. Forecast of the total number of people and households in the region. 2. Based on the future labor force, forecast the number of jobs. 3. Predictions about the location, type, and size of residential and non-residential development projects which would be needed to accommodate the expected growth in households and jobs. 4. The expected growth in households and jobs is allocated to different areas. The models that compose the forecasting system are: Demographic Evolution Model, Employment Model, Real Estate Development Model, Household Location Model, Employment Location Model	
Counties under MPO's Governance: Brazoria, Chambers, Fort Bend, Galveston, Harris, Liberty, Montgomery, Waller	
Baseline Year: 2015	**Population at Baseline Year**: 6.5 million
Horizon Year: 2045	**Population at Horizon Year**: 10.5 million
Forecast Years: 2020, 2025, 2030, 2035, 2040, 2045	
Consultant: -	
Updates produced by: In-house	**Frequency of Updates:** every 2 years
Date of Most Recent Report: November 2017	
Socioeconomic and Demographic (SED) Forecasting Tool: land use model similar to the Urbansim (SAS programming software)	
Zonal Allocation Process (ZAP) Tool: same as SED	
Land Use Model: same as SED	
Staff: 3 modelers, 3 GIS analysts, 1 GIS application developer	
Number of Variables Produced: approx. 20 • Population by 4 race/ethnicity categories (non-Hispanic White, Hispanic White, Black, and Other) • Population by sex • Population by 111 age categories (single year, from 0 through 110) • Employment by 20 2-digit NAICS • Future land use predictions • Household location (3 mile grid) • Employment Location.	
Point of Contact: Pramod Sambidi	**Email:** Pramod.Sambidi@h-gac.com
Links: • https://www.h-gac.com/getmedia/6f706efb-9c6d-4b6a-b3aa-7dc7ad10bd26/read-documentation.pdf	

Atlanta Regional Commission (ARC)	
Overview: ARC uses REMI tools to produce their forecasts. REMI is a structural economic forecasting and policy analysis model. It integrates input output, computable general equilibrium, econometric, and economic geography methodologies. The model is dynamic, and produces forecasts responsive to wage, price, and other economic factors. Base-year forecasts are developed in-house replacing the base-year forecast provided by consultant's model. The production of forecasts using REMI's tools is just the first step of a stepwise approach that includes the ratification of the numbers produced by an advisory committee. Finally, ARC uses a tool purchased by a consultant (TAZ-D) for zonal allocation.	
Counties under MPO's Governance: Barrow, Bartow, Carroll, Cherokee, Clayton, Cobb, Coweta, Dawson, Dekalb, Douglas, Fayette, Forsyth, Fulton, Gwinnett, Hall, Henry, Newton, Paulding, Rockdale, Spalding, Walton	
Baseline Year: 2015	**Population at Baseline Year**: 5.7 million
Horizon Year: 2050	**Population at Horizon Year**: 8.6 million
Forecast Years: 2020, 2025, 2030, 2035, 2040, 2045, 2050	
Consultant: Regional Economic Models, Inc.	
Updates produced by: In-house	**Frequency of Updates:** Every 3-4 years
Date of Most Recent Report: August 2021	
Socioeconomic and Demographic (SED) Forecasting Tool: REMI TransSight	
Zonal Allocation Process (ZAP) Tool: TAZ-D (purchased from consultant)	
Land Use Model: No	
Staff: three full-time and two part-time staff	
Number of Variables Produced: 25 • Employment by 20 sectors • Population by 4 ethnic characteristics • Households	
Point of Contact: Colby Lancelin	**Email:** clancelin@atlantaregional.org
Links: • https://atlantaregional.org/atlanta-region/population-employment-forecasts/ • https://atlantaregional.org/browse/?browse=type&type=data-maps	

North Central Texas Council of Governments (NCTCOG)	
Overview: For socio-economic forecasting, county control totals are obtained from the state demographic forecast and at least two independent private companies. The values are completely reviewed internally. Regional control totals are obtained by analyzing historical trends. County control totals created through this method are used for the validation of the allocation model to the small geographies within the region. For zonal allocation, NCTCOG has developed a method that considers densities and land use. The allocation process also includes local governments review of inputs, assumptions, and outputs that may override the model results. The process includes short-term, based on recent data, and long-term, based on control totals and model allocation, forecasts. For example, in the most recent round, 2005, 2010 and 2015 are observed years, 2016 to 2019 are estimated based on the short-term forecast process, 2030 through 2045 are based on long-term forecast process, and the transition midterm years are between 2020 and 2030. Before finalized, a draft forecast is sent out for review by state, county, and other local governments to suggest changes and provide input.	
Counties under MPO's Governance: Collin, Dallas, Denton, Ellis, Hood, Hunt, Johnson, Kaufman, Parker, Rockwall, Tarrant, Wise	
Baseline Year: 2015	**Population at Baseline Year**: 5.7 million
Horizon Year: 2045	**Population at Horizon Year**: 11.2 million
Forecast Years: 2020, 2025, 2030, 2035, 2040, 2045	
Consultant: two private companies	
Updates produced by: In-house	**Frequency of Updates:** every 4 or 5 years
Date of Most Recent Report: October 2017	
Socioeconomic and Demographic (SED) Forecasting Tool: purchased externally, but reviewed in-house	
Zonal Allocation Process (ZAP) Tool: proprietary model	
Land Use Model: No	
Staff: 2 full-time staff, and 4 part-time	
Number of Variables Produced: 3 • Population • Employment • Households	
Point of Contact: Kathy Yu	**Email:** KYu@nctcog.org
Links: • https://data-nctcoggis.opendata.arcgis.com/documents/nctcog-2045-forecast-city-approximations/explore • https://rdc.dfwmaps.com/MethodologyDocs/NCTCOG%202045%20Forecast%20Description.pdf • https://rdc.dfwmaps.com/PDFs/NCTCOG%202040%20Forecast%20Description.pdf • https://www.nctcog.org/trans/about/staff	

Delaware Valley Regional Planning Commission (DVRPC)	
Overview: DVRPC computed three alternative 2045 population forecasts were calculated for each county based on three separate methods: (i) traditional age- cohort survival model, (ii) redistributing the total 2045 regional population to each of the nine counties based on the county percentages from the adopted 2040 forecasts, and (iii) applying the growth rates between each five-year period from DVRPC's adopted 2040 forecasts. Forecasts for 2020, 2025, 2030, 2035, and 2040 were then calculated by DVRPC, based on the population growth rate predicted over each five-year increment by the Commission's age-cohort survival model. For employment forecasts, DVRPC has traditionally based its long-range forecasts on employment data from the American Association of State Highway and Transportation Officials' (AASHTO) Census Transportation Planning Products (CTPP). In the latest round of forecasts, DVRCP purchased the forecasts from a consultant (IHS Markit). In previous forecasts, they implemented an age-cohort model). The final regional and county controls were obtained after a discussion with all the partners. Zonal allocation is made by UrbanSim's UrbanCanvas.	
Counties under MPO's Governance: Bucks, Chester, Delaware, Montgomery, Philadelphia, Burlington, Camden, Glocester, Mercer	
Baseline Year: 2015	**Population at Baseline Year**: 5.6 million
Horizon Year: 2050	**Population at Horizon Year**: 6.4 million
Forecast Years: 2020, 2025, 2030, 2035, 2040, 2045, 2050	
Consultant: IHS Markit	
Updates produced by: In-house	**Frequency of Updates:** every 4 years
Date of Most Recent Report: July and October 2016	
Socioeconomic and Demographic (SED) Forecasting Tool: -	
Zonal Allocation Process (ZAP) Tool: Urban Canvas	
Land Use Model: UrbanCanvas	
Staff: one full time employee	
Number of Variables Produced: approx. 20 • Population by 5-year increments • Population by sex • Employment	
Point of Contact: Benjamin Gruswitz	**Email:** bgruswitz@dvrpc.org
Links: • https://www.dvrpc.org/Reports/ADR022.pdf • https://www.dvrpc.org/Reports/ADR023.pdf	

Metropolitan Washington Council of Governments (MWCOG)	
Overview: The Cooperative Forecasting program is a peer-reviewed technical process by the local government (county and city) staff. Approximately 16 local governments participate in the consensus building. Forecasts are approved by the Cooperative Forecasting and Data Subcommittee, the Planning Directors Committee, and the COG Board of Directors. The sum of the preliminary local government benchmark projections for total employment, total population and total households are compared to the regional econometric model totals for each of the forecast years. By Planning Directors policy, the 2 sets of projections may vary by no more than 3 percent for each forecast year The forecasting process is an on-going program and every year jurisdictions make a self-assessment and determine (with new transportation facilities or a change in land use) the need to adjust or update their forecast	
Counties under MPO's Governance: City of Frederick, Frederick, Charles, City of Manassas Park, City of Manassas, Prince William, Loudoun, City of Falls Church, City of Fairfax, Fairfax, Prince George's County, City of Gaithersburg, City of Rockville, Montgomery, City of Alexandria, Arlington, District of Columbia	
Baseline Year: 2015	**Population at Baseline Year**: 5.5 million
Horizon Year: 2045	**Population at Horizon Year**: 6.8 million
Forecast Years: 2020, 2025, 2030, 2035, 2040, 2045	
Consultant: -	
Updates produced by: In-house	**Frequency of Updates:** every 5 years - incremental updates from jurisdictions can occur annually upon request
Date of Most Recent Report: November 2017	
Socioeconomic and Demographic (SED) Forecasting Tool: cooperative approach, proprietary econometric models	
Zonal Allocation Process (ZAP) Tool: cooperative approach (done by local jurisdictions)	
Land Use Model: No	
Staff: One coordinator of the Cooperative Forecasting and Data Subcommittee, and four support the committee and technical process	
Number of Variables Produced: approx. 20 • Population by 7 age groups • Employment by 12 categories (sectors) • Households	
Point of Contact: Gregory Goodwin	**Email:** ggoodwin@mwcog.org
Links: • https://www.h-gac.com/getmedia/6f706efb-9c6d-4b6a-b3aa-7dc7ad10bd26/read-documentation.pdf	

Baltimore Regional Transportation Board (BMC)	
Overview: Baltimore Regional Transportation Board does not have a centralized socio-economic and demographics (SED) forecasting methodology. Local agencies do their forecasts independently and submit the results. Local planning agencies that comprise the Cooperative Forecasting Group (CFG) develop their own estimates and projections of population and households based on local comprehensive plans, adopted zoning maps and regulations, and an inventory of available residential holding capacity. These forecasts by small area are submitted to the BMC staff for incorporation into the full round of socio-economic inputs to the travel demand model input variables are updated on an annual basis. Other travel demand model inputs such as the Land Use Classification by Area Type and Total Acreage have already been determined and usually remain as a constant in the forecast set throughout the decade. Although small updates are produced every year, new round of SED forecasts are completed on an as needed basis, usually every new decennial Census.	
Counties under MPO's Governance: Anne Arundel, Baltimore County, Carroll, Harford, Howard	
Baseline Year: 2015	**Population at Baseline Year**: 2.8 million
Horizon Year: 2050	**Population at Horizon Year**: 3.1 million
Forecast Years: 2020, 2025, 2030, 2035, 2040, 2045, 2050	
Consultant: forecast produced by local jurisdictions	
Updates produced by: Local jurisdictions	**Frequency of Updates:** annual "small" updates, and complete rounds of new SED forecasts as needed
Date of Most Recent Report: December 2007	
Socioeconomic and Demographic (SED) Forecasting Tool: Cooperative Forecasting	
Zonal Allocation Process (ZAP) Tool: Cooperative Forecasting	
Land Use Model: -	
Staff: Unkown	
Number of Variables Produced: 5 • Population • Households • Employment • Median household income • School enrollment	
Point of Contact: Shawn Kimberly	**Email:** skimberly@baltometro.org
Links: • https://www.baltometro.org/community/planning-areas/demographic-socioeconomic-forecasting • https://www.baltometro.org/sites/default/files/bmc_documents/general/community/demo-and-socio-forecasting/PrimerSocioEconomic.pdf	

3.4 Lessons Learned from Other MPOs

- There is no one-fits-all approach. Every MPO uses a method that is compatible with their needs and limitations.
- The methods adopted are more related with the resources available (data availability, staff, hierarchical organization in the MPO) than with technical solutions in general.
- All MPOs are striving to improve the adopted methodology to produce the best results possible.
- Hiring consultants is a popular option, but no MPO takes the results produced by the consultants as a given:
    - In some cases, the forecasts produced by the consultants are used as additional information to produce the final forecasts that include inputs from the members of the MPO;
    - The most common case is when MPO hires the consultant for their tools and the MPO staff work along the consultants to get the forecasts. Later, the MPO staff is trained on the tools and can produce updates independently.
- MPOs that use integrated transportation-land use models do their forecasts are part of the modeling process:
    - They model long term transportation decisions such as household location depending on employment location, accessibility, land development, etc.
    - These models require a lot of effort to implement. The data needs are humongous compared to more traditional transportation models, making it prohibitively for some MPOs to adopt this method.
- The most popular method for sociodemographic forecasting is Cohort-Component models:
    - It is relatively inexpensive to apply, provides fairly accurate results, and allows to incorporate external effects based on local knowledge of events that could affect the results.
- Some MPOs use tools that incorporate land use for zonal allocation. This way the control totals are obtained with traditional forecasting models but the allocation to smaller zones takes into consideration the land use distribution of the zone.
- Regarding the tools used, most of the MPOs use a programming language to implement their proprietary models. The most popular are R and Python
- MPOs that hire consultants use their tools. The most popular tools are "REMI Policy Insight" and "UrbanSim".
    - Although the MPOs did not disclose the expenditure, it is known that the license of these tools is expensive compared to more general software that could be used to implement forecasting models (i.e., Excel, R), as they already come with embedded models and the license usually also includes some type of consultancy from the developers including software training programs.
    - The main advantage of these type of tools is that they facilitate the creation of different forecasting scenarios.
    - In the case of UrbanSim, it integrates forecasting as an embedded step of the transportation modeling process.
    - One possible disadvantage is the use of the tool as a "black box" that could limit the interpretation and the incorporation of local expertise into the forecasting.
- None of the MPOs use the results produced by any software or models as a final result. An important part of the modeling processing is gathering the inputs from the committee which is usually composed by representants of local jurisdictions and necessary modifications are made based on their expertise.

4. Recommendations

4.1 Evaluation of Forecasting Methods Found in the Literature

This section presents a qualitative evaluation of the sociodemographic forecasting methods found in the literature review. Six methods are considered for the evaluation:

1. Extrapolative methods
2. Cohort-component models
3. Econometric models
4. Microsimulation
5. Machine Learning
6. Integrated Land Use Models

The methods Averaging and Combining, Housing Led Forecasting, and Disaggregation Techniques were not considered in this part of the evaluation. Averaging and Combining implies that the results are obtained from the application of two or more methods. One example is averaging the results obtained by two methods or more, another possibility is using the lower value between the results of two methods. This condition makes the evaluation of Averaging and Combining unfeasible as it could be a combination of any of the six methods listed above.

Housing Led Forecasting was not considered because it cannot fulfill NYMTC's needs. It estimates the variation in number of households and uses an average household size to convert it into population. Hence, it only allows for computation of population totals and does not allow the estimation of a breakdown by age, gender, etc.

Disaggregation Techniques are already adopted by NYMTC in the zonal allocation step. It requires having control totals that are allocated into smaller areas, and for this reason it cannot be adopted independently. Disaggregation Techniques must be adopted as a second step after counties' forecasts are produced by another method (top-down forecasting approach) or they become unnecessary in the case of the adoption of a technique that produces results at TAZ level directly, in this case values can be aggregated at county level (bottom-up approach).

Table 4 lists the six methods considered and whether they allow for top-down or bottom-up approach. If they allow for bottom-down approach, they can be used to produce forecasts at county level and later be combined with disaggregation technique to produce results at a TAZ level. The reliability of results produced with a bottom-up approach is usually better. One of the reasons that could lead to a better reliability is the quality of the input data that probably has high level of detail to allow for a more disaggregate forecasting in the first place. By default, the level of efforts to implement a method on a more disaggregate scale is higher. It requires more data to cover the heterogeneities among smaller zones and it requires more computational power. A simple example to grasp the added complexity of decreasing the level of aggregation is to model migration among zones. There are 5,418 TAZs among 28 counties in the NYBPM area. Considering migrations among counties would give 756 (28 x 27 = 756) possibilities of migration flows; considering migration among TAZs would generate 29,349,306 possibilities of migration flows.

Table 4: Forecasting methods and their compatibility with top-down or bottom-up approaches

	Top-Down	Bottom-Up
Extrapolative Methods	✔	✔
Cohort-Component Models	✔	✔
Econometric Models	✔	✔
Microsimulation		✔
Machine Learning	✔	✔
Integrated Land Use Models		✔

The six methods were evaluated in five criteria; three related to the effort of implementation of the method—cost, workforce, and data needs—and the other two related to the results produced by the method—flexibility and accuracy. All the categories were assigned a grade from 1 to 5, 5 being the best rated and 1 the poorest. The grades were assigned based on the findings from the literature and MPOs review.

Cost relates to monetary investment necessary to implement the method, including software license, personnel training, etc.; a method that is inexpensive to implement receives a higher score. The information on the actual cost of the methods is not available. It would have to consider the software licenses, eventual training or hiring more staff, specialized equipment, data collection efforts, among others possible expenses. The direct and indirect costs associated with the implementation of any of these methods could vary depending on current conditions or specific requests. Therefore, the evaluation of cost was based on a comparative scale among the methods.

Workforce is directly related to the number of staff needed to produce results with the method; a method that requires less staff receives a higher score. Data needs is directly related to amount of data required to build the method; the less data needed, the highest the score.

Flexibility is probably the most subjective category. It refers if the method allows to easily produce updates and to be easily adapted to account for unforeseen events, for example the impact of the covid-19 pandemic. The flexibility of the method also translates the level of opportunities to add local input from planners and officials into the forecasting process.

Finally, accuracy of results refers to how close to reality the results produced by the method are. This category is a recurring topic in the literature. Various authors research and compare methods to find the most accurate one. The methods with more accuracy are assigned the highest grade.

Table 5: Qualitative evaluation of forecasting methods

	Cost ($)	Workforce	Data Needs	Flexibility	Accuracy of Results
Extrapolative Methods	5	5	5	2	2
Cohort-Component Models	5	4	4	4	4
Econometric Models	2-4	3	3	5	3
Microsimulation	2	3	1	5	3
Machine Learning	2	2	1	5	3
Integrated Land Use Models	1	2	1	5	5

Extrapolative models are the simplest to implement for not requiring a lot of inputs—just historical data is sufficient to implement them. They could be a good alternative for small updates or short-term forecasts but are not recommended for long-term forecasting due to the limited possibility of incorporating the effects of externalities that could impact sociodemographics (i.e., land use changes, economy changes). For this reason, these methods are highly ranked when it comes to resources to implement but received lower scores for the performance of its results.

Cohort-component models are a good compromise in terms of accuracy and necessary resources to implement. These models can be implemented in opensource programming languages such as Python or R, meaning that it is possible to implement them without buying expensive software licenses. In terms of data needs, cohort-component models are reasonable and still produce accurate results. The literature points out cohort-component models as one of the best performances, even when compared to more modern methods that require larger amounts of inputs such as econometric models or simulations.

Econometric models use explanatory variables to predict sociodemographic variables. One of the main necessary efforts is to collect data for the explanatory variables. One advantage is that it gives flexibility to add as many variables are available and could be interesting to shape sociodemographic changes. These variables could include any changes in land use, including new developments, or economic factors of the region such as GDP growth. Another important part of the implementation is the calibration of the models and the selection of the significant variables. Basically, the developers need to test if the added variables increase the explanatory power of the model through statistical procedures. This process makes the implementation of econometric models more resources consuming than cohort-component models or extrapolative methods, and yet literature shows that the results obtained with econometric models are not as accurate as the results obtained with cohort-component models.

Microsimulations model socioeconomics at the individual level. The main advantage is the level of detail of the results that are completely disaggregated, providing outputs at an individual level. Additionally, a custom-made simulation can incorporate the influence of any real-life factors that influence the behavior of individuals, giving it flexibility. However, this level of detail comes with a cost of high data requirements for implementation and potentially the need of more sophisticated hardware to run large metropolitan areas. There is a trade-off: the more elements are included to mimic real-life conditions, the higher the level of efforts to implement the simulation. The process of building a simulation is not straightforward and could lead to months of coding and debugging. Hence, microsimulation received low scored in terms of resources required for implementation.

Machine learning is the cutting-edge research on forecasting. It requires high level of expertise to implement. The current research shows that these models are still lagging in comparison to traditional forecasting models. For this reason, machine learning received average scores when it comes to results. Although this data-driven field has potential, these methods do not seem to be ready for real-life application.

The last method on the table, integrated land use models, is considered the best in accuracy of results. It takes into consideration land use patterns and travel impedance to locate households and employment. The downside of land use models is that they require various inputs that might not be readily available specially in a metro area that is spread through multiple jurisdictions. The effort of collecting and standardizing the input data could make the implementation of integrated land use models prohibitively costly. In addition, it requires purchasing software licenses or hiring consultancy from developers of integrated land use models tools.

Another point to be taken into consideration is the compatibility of the methods with the NYBPM. The NYBPM is an agent-based transportation model that requires sociodemographic at TAZ level to later generate a synthetic population. In that sense, any of the core forecasting methods are compatible with the NYBPM. To ensure compatibility, the forecasting results must be produced at TAZ level and that could be obtained with all the methods except the integrated transportation land-use model. As the main objective of an integrated transportation land-use model is to model transportation, it would be a substitute of the NYBPM in place.

Table 6: Summary of main advantages and disadvantages of the listed methods.

Method	Advantages	Disadvantages
Extrapolative Methods	Easy to implement	Based on past trends only
Cohort-Component Models	Good "cost/benefit"	Limited consideration of external factors (e.g., land use)
Econometric Models	Incorporation of spatial and economic variables	Large data requirements and complicated calibration process
Microsimulation	High level of detail of outputs	Large data requirements and computational requirements
Machine Learning	Data-oriented nature	High level of expertise required and results are not necessarily as acurate as simpler methods
Integrated Land Use Models	Accuracy of results and incomporation of land use-transportation cycle	Large data requirements and specialized software

In essence, cohort-component models are an adequate method for NYMTC. They produce adequate results and are a good cost/benefit solution for forecasting. In addition, the NYMTC team is already familiar with the methodology and it is fully compatible with the NYBPM . The literature covers variations of cohort-components models. The error difference among them is low, however there could be opportunities to enhance the current adopted methodology by using a different variation of the method. For instance, results in the literature show that constrained cohort-component models provide a better accuracy than the constrained bi-regional cohort component model. In this case, the values produced for the forecast are capped by an extrapolative method.

4.2 Evaluation of Zonal Allocation Methods Found in the Literature

The zonal allocation method is referred as disaggregation techniques in the literature. It is relevant when there is data available for larger areas and there is a need of a more granulated level of detail. It is not a topic fully explored in the literature and revolves as a subtopic of forecasting, often called “downscaling” or “top-down” forecasting.

The technique is based on distributing the values to smaller zones. The most basic approach is to weigh the distribution by land area; however, it is possible to use other variables to make the distribution more realistic such as urban area, number of buildings, type of land uses, etc.

To enhance the zonal allocation process, NYMTC could explore one of the databases that describe buildings or parcels in the metropolitan area. Some suggestions are as following:

- OpenStreetMaps Buildings: is a crowdsourced database that includes street networks, landmarks, parks, buildings, etc.. (https://wiki.openstreetmap.org/wiki/Buildings). For buildings, it provides a classification system regarding the use of the building: accommodation, commercial, religious, civic/amenity, agricultural, sports, storage, cars, technical, and others. The classification also includes more specific subcategories, e.g., apartments. Other attributes of the buildings are also available, such as footprint area and height. It is possible to download polygon shapefiles with their building database through the python package OSMnx (https://osmnx.readthedocs.io/en/stable/). The main advantage of OpenStreetMaps tool is that it is opensource and there are no geographical limitations. However, OpenStreetMaps is crowdsourced; the information needed might not be necessarily available or standardized. For instance, although the building dataset is comprehensive, many buildings lack tags regarding its classification.
- New York City Zoning Information: the city provides information on its zoning system in the form of shapefiles (https://zola.planning.nyc.gov/about/#9.72/40.7125/-73.9022; https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-gis-zoning.page). The areas are larger than parcels and refer to the types of activities that are allowed in the zone, e.g., residential, commercial. Hence, the information is not as insightful as most of the zones are some types of mixed land use that allows a combination of residential dwellings and commercial establishments. The data is official and reliable but is limited to NYC’s boundaries.
  New York State Tax Parcels: NYS collects tax parcels information from counties and conveniently assembles a unified database (http://gis.ny.gov/parcels/). The website contains tax parcels polygon shapefiles only for some counties that make them available. Alternatively, there is a comprehensive shapefile with tax parcels centroids for the entire state. The tax parcels have information on property type. Note that the NYS classification system (https://www.tax.ny.gov/research/property/assess/manuals/prclas.htm) differs from NYC’s, which is called “building style codes” (https://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html). New Jersey Tax Parcels: New Jersey also keeps a database of statewide tax parcels (https://www.arcgis.com/home/item.html?id=406cf6860390467d9f328ed19daa359d).
  The New Jersey MOD-IV User Manual provides an in-depth explanation on the fields associated for each tax parcel, including property classification codes that dictate the type of use of the parcel (https://www.state.nj.us/treasury/taxation/pdf/lpt/modIVmanual.pdf).
- Business by Industry Sector and ZIP Code: the United States Census Bureau provides information on businesses and employment at ZIP Code level (https://www.census.gov/data/datasets/2019/econ/cbp/2019-cbp.html). Business counts and employment are described on 2-digit level NAICS or, in some cases, 3-digit

level. The database is in table format but can be used in GIS if combined with a ZIP Code shapefile layer. The main advantage of this database is that it provides NAICS of business that is not informed in tax parcels, but it has a much higher level of aggregation.

All the databases have advantages and disadvantages, and it could be that a better solution is to mix and match information across them. The incorporation of any land use related information to the zonal allocation process is not a straightforward task and could potentially require multiple attempts using different datasets and variables. However, the incorporation of this information could be highly beneficial to the zonal allocation process.

From the suggested data sources, tax parcels are a good point of start. The information is comprehensive, comes from official sources and is standardized. The zonal allocation could be done by weighing parcel counts of a certain type or parcel area.

4.3 Final Remarks

When comparing to the methodologies adopted by other MPOs, the forecasting methodology currently adopted by NYMTC is robust and it is not outdated. The survey showed a current trend of MPOs migrating to integrated land use-transportation models, however most of them continue using other forms of transportation modeling. In addition, each MPO has unique characteristics, challenges, and needs that must be taken into consideration before making any direct comparisons between methodologies adopted in different MPOs. For starters, New York City is the largest metropolitan area in the country; the only MPO with a comparable population is the Southern California Association of Governments (SCAG). In addition, NYMTC has the added challenge to being in a tri-state area meaning that the challenge of coordination among agencies and homogeneity of the input data is potentially larger than any other MPO whose metropolitan area falls within the limits of a single state.

The search for forecasting models in the literature showed that the state-of-the-art methods (simulations and machine learning approaches) are still crude. Besides requiring a high level of expertise to implement, they are outperformed by more traditional methods such as cohort-component models. From the newer methods, the only method that seems worthy of attention in terms of quality of results are integrated transportation land use models. Nonetheless, the effort of researching and keeping up with the current trends is necessary to maintain the relevance of the implemented methods and the quality of the results produced.

With that in mind that are a few recommendations that the NYMTC team could pursue:

- Migrate the current Excel based models to a more friendly coding language (e.g., python) to improve the readability and the maintenance of the models.
- Although the current zonal allocation process already incorporates land use features to weight the distribution from counties to TAZs, there are other datasets that could be explored to potentially enhance the process.
- Explore variations of the cohort-component models, or even simple combinations with other methods that are reported to produce more accurate results in the literature, for example the constrained bi-regional cohort-component model.
- Continue the effort of always in touch with the current trends in forecasting and modeling. We live in a time of rapid advent of technology and digitalization; the limitations that prevent the implementation of a method today could be lifted sooner than we expect.

References

Baker, J., Alcántara, A., Ruan, X., Watkins, K., & Vasan, S. (2014). Spatial weighting improves accuracy in small-area demographic forecasts of urban census tract populations. *Journal of Population Research*, *31*(4), 345-359.

Ballas, D., Clarke, G. P., & Wiemers, E. (2005). Building a dynamic spatial microsimulation model for Ireland. *Population, Space and Place*, *11*(3), 157-172.

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451-468.

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International journal of forecasting*, *22*(3), 547-581.

Breidenbach, P., Kaeding, M., & Schaffner, S. (2019). Population projection for Germany 2015–2050 on grid level (RWI-GEO-GRID-POP-Forecast). *Jahrbücher für Nationalökonomie und Statistik*, *239*(4), 733-745.

Chi, G., & Wang, D. (2017). Small-area population forecasting: A geographically weighted regression approach. In *The frontiers of applied demography* (pp. 449-471). Springer, Cham.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, *5*(4), 559-583.

Goodwin, P. (2009). New evidence on the value of combining forecasts. *Foresight: The International Journal of Applied Forecasting*, (12), 33-35.

Hamilton, C. H., & Perry, J. (1962). A short method for projecting population by age from one decennial census to another. *Social Forces*, *41*(2), 163-170.

Hauer, M. E., Evans, J. M., & Alexander, C. R. (2015). Sea-level rise and sub-county population projections in coastal Georgia. *Population and Environment*, *37*(1), 44-62.

Hunt, J. D., & Abraham, J. E. (2005). Design and implementation of PECAS: A generalized system for allocating economic production, exchange and consumption quantities. In Integrated Land-Use and Transportation Models (pp. 253-273). Emerald Group Publishing Limited.

Isserman, A. M. (1977). The accuracy of population projections for subcounty areas. *Journal of the american Institute of Planners*, *43*(3), 247-259.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one*, *13*(3), e0194889.

Marois, G., & Bélanger, A. (2014). Microsimulation Model Projecting Small Area Populations Using Contextual Variables: An Application to the Montreal Metropolitan Area, 2006-2031. *International Journal of Microsimulation*, *7*(1), 158-193.

Mazzuco, S., & Keilman, N. (2020). *Developments in demographic forecasting* (p. 258). Springer Nature.

Raftery, A. E., Li, N., Ševčíková, H., Gerland, P., & Heilig, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*, *109*(35), 13915-13921.

Rayer, S. (2008). Population forecast errors: A primer for planners. *Journal of Planning Education and Research*, *27*(4), 417-430.

Rayer, S., & Smith, S. K. (2010). Factors affecting the accuracy of subcounty population forecasts. *Journal of Planning Education and Research*, *30*(2), 147-161.

Šimbera, J. (2020). Neighborhood features in geospatial machine learning: The case of population disaggregation. *Cartography and Geographic Information Science*, *47*(1), 79-94.

Simpson, L. (2017). Integrated local demographic forecasts constrained by the supply of housing or jobs: Practice in the UK. In *The Frontiers of applied demography* (pp. 329-350). Springer, Cham.

Smith, S. K. (1987). Tests of forecast accuracy and bias for county population projections. *Journal of the American Statistical Association*, *82*(400), 991-1003.

Swanson, D. A., Schlottmann, A., & Schmidt, B. (2010). Forecasting the population of census tracts by age and sex: An example of the Hamilton–Perry method in action. *Population Research and Policy Review*, *29*(1), 47-63.

Waddell, P. (2002). UrbanSim: Modeling urban development for land use, transportation, and environmental planning. Journal of the American planning association, 68(3), 297-314.

Waddell, P., Boeing, G., Gardner, M., & Porter, E. (2018). An integrated pipeline architecture for modeling urban land use, travel demand, and traffic assignment. arXiv preprint arXiv:1802.09335.

Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., ... & Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences*, *115*(14), 3529-3537.

Weatherby, C. (1995, February). Summary of workshop observations and recommendations. In Land Use Modeling Conference ProceedingsTexas Transportation InstituteFederal Highway AdministrationFederal Transit AdministrationOffice of the Secretary of TransportationEnvironmental Protection AgencyDepartment of Energy (No. DOT-T-96-09).

Wegener M, Fürst F (1999) Land-use transport interaction: state of the art. Berichte aus dem Institut für Raumplanung 46. Institute of Spatial Planning, University of Dortmund, Dortmund

Wegener, M. (2021). Land-use transport interaction models. Handbook of regional science, 229-246.

White, H. R. (1954). Empirical study of the accuracy of selected methods of projecting state populations. *Journal of the American Statistical Association*, *49*(267), 480-498.

Wilson, T. (2015). New evaluations of simple models for small area population forecasts. *Population, Space and Place*, *21*(4), 335-353.

Wilson, T. (2016). Evaluation of alternative cohort-component models for local area population forecasts. Population Research and Policy Review, 35(2), 241-261.

Wilson, T., Grossman, I., Alexander, M., Rees, P., & Temple, J. (2021). Methods for small area population forecasts: state-of-the-art and research needs. *Population research and policy review*, 1-34.