Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6672961, 15 pages
https://doi.org/10.1155/2021/6672961

Research Article

Revisiting the Benefits of Combining Data of a Different Nature: Strategic Forecasting of New Mode Alternatives

Luis A. Guzman,[1] **Julian Arellana**,[2] **Victor Cantillo-García**,[1] **and Juan de Dios Ortúzar**[3]

[1]*Department of Civil and Environmental Engineering, Universidad de Los Andes, Bogotá, Colombia*
[2]*Department of Civil and Environmental Engineering, Universidad Del Norte, Barranquilla, Colombia*
[3]*Department of Transport Engineering and Logistics, Instituto Sistemas Complejos de Ingeniería (ISCI), BRT+ Centre of Excellence, Pontificia Universidad Católica de Chile, Santiago, Chile*

Correspondence should be addressed to Luis A. Guzman; la.guzman@uniandes.edu.co

Received 1 December 2020; Accepted 14 July 2021; Published 23 July 2021

Academic Editor: Michela Le Pira

We revisit the practice of combining revealed (RP) and stated preference (SP) data (i.e., the data enrichment, DE, paradigm) in discrete choice models using secondary data obtained from emerging sources; these facilitate access to massive information about travel choices and can be used to improve transport models. Even though the benefits of the DE paradigm have been known for years, there is a large gap between the state of practice and the state of the art, particularly in Global South countries (but also in many industrialized nations). We use a SP dataset considering two new transport alternatives (train and metro) and a RP dataset based on a large mobility survey in Bogotá, Colombia, complemented with fairly precise level-of-service data obtained using GIS utilities and the Distance Matrix API by Google. Our results allow us to discuss good practice, identify barriers and challenges to the paradigm's application, and draw recommendations for forecasting the demand for new alternatives using joint RP and SP data.

1. Introduction

Emerging technologies and telematics offer tools to access and collect new good-quality data that might improve transport models. Many of these tools refer to nondisruptive mechanisms that capture travel paths generated by electronic devices, such as smartphones, geographic positioning (GPS), transactional data, and sensors [1], which are usually characterized by a large number of near real-time observations (*big data*). Moreover, web-based experiments are an alternative to gather active data, achieving representative samples at a lower cost compared to traditional data collection methods such as face-to-face and telephone surveys. On the other hand, data disparity is particularly challenging because it is necessary to group and homogenize information of a different nature.

Revealed (RP) and stated preference (SP) data are two types of data commonly used in transport research, particularly for the estimation of discrete choice models to analyze travel behaviour. RP data refer to actual choices made by individuals in a specific context. This information is useful to understand preferences in existing markets, but it is generally inappropriate to try and model preferences in the case of new alternatives. SP data include a range of techniques designed for gathering information about preferences in hypothetical scenarios representing a market context specifically designed by the modeller. SP data are useful to understand individual behaviour in the presence of new alternatives, grasp the importance of difficult-to-measure variables, and estimate substitution patterns among attributes [2].

Most passive emerging mobility data can be considered RP information. For instance, public transport smartcard validations, GPS records, or speed sensors enable to capture the current user's behaviour. In contrast, depending on the design, active online surveys allow to collect of both RP and

SP data. These are rich sources to gather information about mobility patterns, preferences, choices, and operational indicators that can be used to improve the capabilities of transport models. However, there are some challenges to their widespread use, especially in the Global South (The term Global South was coined some years ago in the context of the North-South divide. It is made up of Africa, Latin America and the Caribbean, Pacific Islands, and the developing countries in Asia, including the Middle East. It is generally seen as home to Brazil, India, Indonesia, and China, which, along with Nigeria and Mexico, are the largest Southern states in terms of land area and population (source Wikipedia).) context where data availability and quality are issues to address. Also, big data from passive sources lack the multidimensional disaggregate information necessary to construct robust transport models, and web-based surveys require a comprehensive design to reduce the risk of response bias due to the self-response nature of the experiment. In this context, how to combine these different data sources remains a challenge [3, 4]. Therefore, improvement of transport model implementation in Global South cities should be approached with care. Preferences and travel choices are likely to be significantly different, and proper adjustments are needed.

In this context, we formulate a mode choice model for the city-region of Bogotá, considering the combination of RP data built from a large household mobility survey, GIS utilities and travel time data from Google, and an existing SP survey. Furthermore, a special requirement for the model was that it should eventually be integrated into a Land-Use Interaction (LUTI) model, as part of a project to support local and regional governments' decision-making processes. In this context, the capacity to forecast new alternatives is crucial. The paper aims to provide some methodological insights into joining information from emerging data sources with SP data based on the well-established data enrichment paradigm [2], by means of a case study.

The rest of the paper is organized as follows. Sections two and three review the current literature and the benefits of combining data of a different nature. Section four presents a brief description of the case study. Section five discusses the data enrichment approach and reviews the theoretical framework for estimating discrete choice models with mixed RP-SP data. Section six summarizes our modelling approach, and specifications, while Sections seven and eight include the results, discussion, and conclusions.

2. Literature Review: Benefits of Mixing SP-RP Data

The spatial and temporal detail of information from emerging data sources can be used to enrich conventional transport data when modelling transport behaviour [4–7]. On the other hand, social media and web platforms might enhance traditional data collection methods [8]. Some applications show that modern passive data from smartphones and smartcards have been successfully used in the estimation of disaggregate origin-destination matrices [9, 10]. Phone call records can be used to analyze the relations between social networks and travel behaviour [11], while the combination of GPS data, location of points of interests, and street view imagery have a great potential in the analysis and prediction of active modes' preferences [12, 13].

RP data are mainly collected through mobility surveys; these require high-quality and precision standards when their intended use is to model discrete choices [14], leading to high data collection costs. Unfortunately, in Global South cities, these costs may not fit the resources available for the studies and the analysts tend to redefine the project objectives or settle for less detailed information. On the contrary, SP data collection is usually cheaper because there is no need to measure in detail the travel attributes of all competing alternatives and sadly because the careful design steps required to make it a trustworthy data source are usually ignored ([14], Section 3). In fact, and paradoxically, the tradition for SP studies in the Global South has been to allocate scarce resources to the experimental design stage, focusing mainly on the in-field data collection process, leading to less-than-ideal quality data.

Money and time constraints are not the only issues about RP and SP data when used as stand-alone sources. Usually, RP data do not offer enough attribute-level variability for specifying robust models for evaluation and forecasting. Also, observed choices may be dominated by a few factors, making it hard to assess the relevance of secondary or latent attributes. Finally, it is very hard to evaluate the implementation of new alternatives and policies with RP data. However, even though some SP data features allow to overcome some of the disadvantages of RP data, the approach also has some downsides [2]. RP data may be considered less flexible since they provide information regarding specific market conditions at a specific time. In contrast, SP data may lack realism due to the artificiality of the choice scenarios or, in many cases, a bad design. Also, SP data can have errors due to respondent fatigue, policy response bias (in the case of new alternatives), or self-selectivity bias [15]. This is, sadly, a recurring case in the Global South, because authorities refuse to share detailed information, or there is not adequate expertise to properly design and conduct SP studies.

The methods and benefits associated with combining RP and SP data have been widely reviewed in the academic literature [2, 16–20]. Besides, improvements in computational capabilities and the availability of extremely powerful and sophisticated free software [21, 22] have fostered an increase in the number of applications of this nature in different contexts in the last years. Combining data sources has proven to be adequate to analyze the disruptions of new alternatives [23, 24] and to evaluate user preferences and policy scenarios [25, 26]. Other applications can be found in the context of air transport [27], and the joint modelling of mode and departure time [28, 29], among others. The combination of data sources to estimate discrete choice models has also been applied to evaluate the preference for food products [30].

There is an academic consensus that both RP and SP data have complementary strengths, so it is not surprising that combining them may allow overcoming the difficulties of

using either type of data alone, enhancing the modelling process. This is especially relevant considering the increasing availability of new data sources with massive information potential. Nevertheless, joint RP-SP models are not common in practical real-world projects. Additionally, standard practice to estimate and use mixed data models is inadequate, especially in less industrialised countries. Then, it appears convenient to re-examine the use of this powerful technique. Furthermore, there are still some methodological questions associated with using pooled data models, including which set of parameters should be used when forecasting, and how to specify the model structure to guarantee that the data enrichment process yields unbiased results [31].

3. Combining RP and SP Data in Discrete Choice Models

The estimation of discrete choice models is based on random utility theory. This theory states that individuals associate a utility to each available alternative and then choose the option with the highest utility [32]. The modeller, who is an observer, assumes that the utility of alternative i for individual q (U_{iq}) is equal to a systematic or measurable utility (V_{iq}) and a random component (ε_{iq}) capturing unobserved preferences and any measurement errors, such that:

$$U_{iq} = V_{iq} + \varepsilon_{iq}. \tag{1}$$

The systematic utility is usually taken as a linear combination of attributes of the alternatives i and characteristics of the individuals q (X_{iq}), each weighted by parameters (β_{iq}) to be estimated (marginal utilities), including a set of alternative specific constants ASC_i. Depending on the structure of the error components, diverse model specifications can be derived. If the error is assumed to be independent and identically distributed (IID) Extreme Value Type I, the Multinomial Logit (MNL) model is derived and the probability of an individual choosing an alternative is given by equation (2), where λ is a scale parameter associated with the unknown standard deviation of the residuals ε_{iq} ([14], Section 7):

$$P_{iq} = \frac{e^{\lambda V_{iq}}}{\sum_{i=1}^{i} e^{\lambda V_{jq}}}. \tag{2}$$

Other popular specifications are the Nested Logit (NL) model and the Mixed Logit (ML) model [33]. These complementary specifications allow relaxing the less realistic assumptions of the MNL (independence and homoscedasticity). In fact, the ML model permits to specify a general covariance structure allowing to treat random taste heterogeneity ([14], Section 7). This formulation is suitable to deal with panel data, by varying the utility across individuals according to a density function $f(\beta \mid \theta)$ conditioned on population parameters θ, but fixing it over the repeated observations of each individual. In this case, the unconditional probability of choosing an alternative by an individual is given by equation (3), where P_{iqk} is the same as in equation (2) but for each repeated choice k. This multidimensional integral can be estimated using simulated maximum likelihood methods [33].

$$P_{iq} = \int \prod_{k} P_{iqk} f(\beta|\theta) \mathrm{d}\beta. \tag{3}$$

The combination of RP and SP data is grounded in the data *enrichment (DE) paradigm* [2]. RP data bring information related to the equilibrium and tradeoff in a particular context, while the SP data help to expand the information regarding tradeoffs in new scenarios. In other words, the objective of combining RP and SP data is to overcome the disadvantage of each type of information. The benefits of this procedure are improvements in the efficiency of preference estimation, bias correction, and identification of preferences for new alternatives [17].

If the DE paradigm fully applies, all attributes of the utility functions of the mixed model, excluding the ASC, should have the same parameters (i.e., all parameters are common). However, typically this is not the case. In this situation, we have *partial data enrichment*. Just one common parameter is sufficient to combine the data from various sources and the remaining coefficients are specific to each data source. Parameters might not be common due to many reasons, including measurement errors, correlations between attributes, or limited variability [2].

Since the standard deviations of the random components of the utility functions in the RP and SP datasets are likely to differ (the data have different measurement errors), a scale parameter μ (equal to the ratio of the scale parameters corresponding to each dataset in isolation, see equation (2)) needs to be estimated. Best practice recommends normalizing the scale factor of the RP data to one, so μ is effectively equal to the scale parameter of the SP dataset ([14], Section 8). This framework allows combining not only two but N sources of information, where $N-1$ of these should be scaled with respect to a reference fixed to one.

In the context of the partial *DE paradigm*, the modeller should apply the following methodology to assess which attributes are common to both domains ([14], Section 8.7):

(1) Estimate models using each dataset alone and obtain the parameters of their respective utility functions.

(2) As the coefficient of a given attribute in the SP set should be equal to the coefficient of this same attribute in the RP set multiplied by the scale factor μ, graph the estimated parameters in both environments on a scatterplot, and expect the points to fall in an elliptical region close to the line that passes through the origin with slope equal to μ.

(3) The parameters falling relatively away from this elliptical zone might not be treated as common (as their difference is real, not just a scaling issue).

(4) In some cases, nonsignificant parameters falling far from the line might still be considered as common (i.e., if their value is fixed, the model does not change). For this reason, the DE paradigm should be tested using a likelihood ratio test (equation (4)), where LogLikelihood refers to the log-likelihood at

convergence of the three models considered (i.e., the joint RP-SP model, the RP-alone model, and the SP-alone model), and LR distributes asymptotically χ^2 with degrees of freedom equal to the number of parameters assumed to be common minus one. The null hypothesis is that the common utility parameters are equal [2].

$$\mathrm{LR} = -2\left[\mathrm{LogLikelihood}_{\mathrm{Joint\,RP-SP}} - \mathrm{LogLikelihood}_{\mathrm{RP}} - \mathrm{LogLikelihood}_{\mathrm{SP}}\right]. \quad (4)$$

4. The Case of Bogotá

Bogotá, the capital city of Colombia, has an urban area of around 380 km^2, with a population of 7.42 million people in 2018. The population distribution is heterogeneous through the territory, resulting in some inequities in access to opportunities [34]. Housing and economic constraints of low-income households have historically led to informal settlements (some later regularised), mainly in the southern and western peripheries, away from the main employment centers [35]. The densest areas in the city, which can reach average values up to 56,000 inhabitants/km^2, are associated with the poorest neighborhoods and lowest socioeconomic strata (SES) (A housing classification system (into six categories, according to its physical characteristics), commonly associated with income levels in Colombia [36]. SES 1 and 2 are identified with poverty and are mainly located in the south. In the centre, middle classes predominate, while higher SES households, considered rich, are mainly located in the northeast of Bogotá).

The average household size in Bogotá was 3.44 in 2015. Most daily trips in the city were made by public transport (39%) or active modes (35%). Specifically, the modal share consisted of 31% walking trips, making it the most used transport mode. Bogotá's land-use patterns involve long commuting trips, forcing low-income households (low-SES) to make frequent complex multimodal trips, usually including long travel distances, causing unequal accessibility levels; therefore, tradeoffs between travel cost and travel time are relevant for these segments [37].

For the last two decades, Bogotá has embarked on public transport reforms aimed at improving and formalizing its bus services. These efforts were initiated with the implementation of a BRT system in the year 2000 [38] and also with the implementation of an integrated public transport system (SITP by its acronym in Spanish). Recent policy interventions consider the implementation of a regional rail system, including the construction of the first metro line connecting the southwest with the central business district. These services should start operating within the current decade. Also, a cable car line was implemented in 2018 to provide an accessibility solution to a southern poor area. Lastly, important efforts to expand cycling infrastructure have been evident in the recent past [39].

Although public transport use remains high despite increases in motorization and shifts in demand towards walking and cycling, Bogotá faces several challenges regarding mobility [40]. The modal share of public transport has lost participation (from 57% in 2005 to 39% in 2015) but still serves a big proportion of the travel demand in the city. Of the current 14.9 million daily trips in Bogotá, 5.8 million are made by public transport. If we focus on motorized trips, the modal share of public transport reaches 61% of daily travel demand.

Under this context, an update of the city's Land-Use and Transport model [41] is currently in progress. This model is a strategic model designed to simulate the dynamic interaction processes between land-use and transport at an aggregate level in Bogotá and its regions. The model is being updated to incorporate new areas and new transport modes such as metro, regional trains, and active transport. Thus, the combined RP-SP model in this paper will be a crucial part of the new strategic Land-Use and Transport model and used to simulate the passenger travel demand on working days in dynamic interaction with activity location models.

5. Data

For the analysis, we used two information sources. First, an online SP survey was conducted during 2018; it was designed by a local consultant firm contracted by the local authorities in Bogotá. Second, an RP databank was constructed by the authors, based on public trip information from the 2015 Bogotá Mobility Survey. We obtained trip characteristics for the nonchosen alternatives from the Distance Matrix API of Google (https://cloud.google.com/maps-platform). The following sections present a more detailed description of each dataset.

5.1. Stated Preference. The SP data component was an online survey with two sections. The first included a series of questions regarding individual socioeconomic data and trip characteristics. The socioeconomic attributes included age, sex, income, education level, occupation, household size, vehicle ownership, and SES. Trip characteristics included information about trip origin and destination, trip purpose, departure time, travel time, cost, and modes used in each travel leg.

The second section of the survey was a choice experiment where each respondent faced eight hypothetical scenarios where they had to choose between a set of transport modes including car, motorcycle, bus, BRT (*Transmilenio*), bicycle, regional train, and metro. All alternatives were described by only two attributes: travel time and cost. Car and motorcycle were available only if the respondent reported vehicle ownership in the first section. Train and metro are new alternatives; that is, they are not available at present in the

city but are under design and expected to start operations within the next decade.

The sample size was estimated assuming a 5% error and a confidence interval of 90%, considering a stratified sample with four segments of the population according to SES (SES 1 and 2, SES 3 to 5, and SES 6). The final sample size was 1,930 individuals with a total of 15,041 observations in the SP dataset (Not everyone completed the eight choice scenarios.). The experimental design was allowed to vary according to trip length (i.e., short, medium, and long) and transport mode. The online survey tool was programmed to calculate travel distances from the origin and destination reported in the first section, and then to present the SP choice situations customized on the reported trip length. Trips less than 4.5 km long were considered short trips, while those longer than 11.5 km were defined as long-distance trips (those between 4.5 km and 11.5 km were considered medium distance trips). The survey instructions asked participants to consider travel time and cost as representatives of the conditions for the entire trip of each transport mode in the experiment. Table 1 presents the attribute levels for each version of the survey.

5.2. Revealed Preference. The RP databank was developed to complement the SP dataset and profit from the benefits of combining both sources of information. We were interested in estimating a model accounting for current transport modes and new transport alternatives in the city. As an important mode (walking) was excluded from the SP component, applying a model estimated on SP data only could lead to biased estimations of the potential mode choices in a real context. Around 36.5% of the total daily trips and 24.7% of the trips longer than 15 min in Bogotá are made on foot. Furthermore, Table 2 shows that train and metro represented 62.3% of the total choices in the SP dataset. This aggregate modal share suggests that SP responses may be biased towards these new modes, so using the SP data alone might not be appropriate.

The RP dataset was constructed in two stages. First, we selected a random sample of trips by zone (The study area was divided into 148 analysis zones) from the 2015 mobility survey of Bogotá. The number of sampled trips by zone was proportional to the total number of trips originating in each zone, so the final sample follows a similar spatial distribution of origins to that observed in the mobility survey. Then, we joined the trip information, the household characteristics, and the individuals' socioeconomic attributes from the different components of the survey. This procedure allowed us to obtain a sample representing the modal shares similar to the population.

From the 2015 Mobility Survey, we compiled the data and chose the mode for each individual. Then, we computed the trip attributes (i.e., cost and time) for the available alternatives. Given that the dataset only included information regarding the chosen mode, our first approximation to impute trip attributes for the nonused alternatives was to calculate the mean values reported in the survey by mode between each origin and destination at the zone level.

Table 1: Travel time and cost levels in the SP experiment.

Trip length	Short		Medium		Long	
Attribute	Time	Cost	Time	Cost	Time	Cost
Motorcycle	10	1.00	25	2.50	40	3.90
	15	1.40	40	3.50	65	5.60
	20	1.80	55	4.60	90	7.30
Car	20	2.20	20	5.70	45	9.10
	30	3.20	30	8.10	75	13.00
	20	4.20	50	10.50	105	16.90
Bus	25	1.40	55	1.40	60	1.40
	40	2.00	40	2.00	100	2.00
	55	2.60	70	2.60	140	2.60
BRT	20	1.60	40	1.60	40	1.60
	30	2.30	30	2.30	70	2.30
	40	3.00	50	3.00	100	3.00
Metro	20	1.60	40	1.60	35	1.60
	30	2.30	25	2.30	55	2.30
	40	3.00	45	3.00	75	3.00
Train	20	1.60	40	1.60	40	1.60
	30	2.30	25	2.30	65	3.00
	40	3.00	45	3.00	90	3.60
Bicycle	20		40		75	
	30	—	45	—	75	—
	40		75		105	

Units: time in minutes and cost in thousand Colombian pesos (COP, at the time of the survey; 1 US$ = 2,956 COP).

Unfortunately, this turned out to be futile since the low variability of the data did not allow to generate robust enough travel times and costs matrices by mode (The seminal paper by Daly and Ortúzar [42] reported that while using aggregate level of service, data at the zonal level was inevitable in the case of destination choice modelling, and it was essential to use disaggregate data, measured at the individual level (i.e., subject to less measurement error), in the case of estimating appropriate mode choice models. Ribeiro et al. [43] confirmed previous finding showing better transport demand model estimates when using GPS measurements instead of self-reported or simulated travel time values). Therefore, we followed a different approach, where attributes of all available alternatives were estimated with the following procedure:

(1) We retrieved travel times and distances for car and public transport from Google's Distance Matrix API, considering the departure time and the georeferenced origin and destination of each trip from the mobility survey. Motorcycle travel time was equal to the travel time by car multiplied by the ratio of their average travel speeds according to the mobility survey results. We estimated walking and bicycle distances using GIS-based shortest path algorithms. Then, we obtained travel times using the average speeds for these modes from the mobility survey.

(2) We obtained travel costs for car and motorcycle considering the distance and the average operational costs (USD 0.20 and 0.05 per km, respectively) used in the socioeconomic evaluation of the first metro

Table 2: Sample share of observed choices by dataset.

Mode	SP			RP		
	Times available	Percentage chosen overall	Percentage chosen when available	Times available	Percentage chosen overall	Percentage chosen when available
Motorcycle	2,355	2.4	15.3	864	5.1	36.7
Car	882	5.4	10.3	1,851	14.8	49.7
BRT	15,041	9.6	9.6	2,675	17.5	40.8
Bus	15,041	9.2	9.2	4,119	31.4	47.4
Train	15,041	17.1	17.1	—	—	—
Metro	15,041	45.2	45.2	—	—	—
Bicycle	15,041	11.2	11.2	2,239	3.1	8.7
Walking	—	—	—	3,101	28.1	56.3

line in Bogotá. In the case of bus and BRT, the cost corresponded to the official fares reported by the operators. Walking and bicycle were assumed to have no cost.

Regarding availability, car, motorcycle, and bicycle were considered available if the individuals reported their ownership in the corresponding section of the mobility survey. Walking was only permissible for trips shorter than 5 km, and bicycle for trips shorter than 10 km. We defined these thresholds by analyzing decay functions using scatterplots of the number of trips by distance for each mode.

The availability of bus and BRT needed some additional considerations since the Distance Matrix API provides the information aggregated as public transport in general, so it was not possible to disaggregate it into these two alternatives. Therefore, if the chosen transport mode for a trip was bus (or BRT) according to the mobility survey, this mode was the only public transport alternative available, and the other was not. If the mode chosen for the trip was different than public transport, BRT was assumed to be available if the distance to a BRT station from the origin was less than 600 m. Bus was considered to be always available since that subsystem covers the study area thoroughly.

Table 3 shows that the sample size of the RP component was 6,221 individuals. There are some differences between the distribution of the attributes on the SP and RP components. The SP dataset mainly comprises young adults of medium-high income and tertiary studies, while the RP data reflect the conditions of Bogotá better, with more significant participation of low-income individuals. Note that all attributes were coded as dummy variables to explore possible heterogeneity according to population segments.

6. Model Formulation, Estimation, and Use

Figure 1 describes the procedure followed to formulate, estimate, and use a combined RP-SP model. This section also presents the methodology designed to make a forecast with the model in the framework of a strategic Land-Use and Transport Interaction model in Bogotá.

The modelling process started with the construction and processing of the databanks. We coded the socioeconomic attributes as dummy variables to test for the presence of preference heterogeneity through interactions of these with the ASC, travel time, and cost attributes (i.e., systematic taste variations). We then estimated separate models for the RP and SP components, checked which attributes could be taken as common to both environments, and specified the combined RP-SP model following the procedure described in the previous section. We estimated the models using the Apollo package [22], available in the *R* software.

The models were assessed considering the expected parameter signs, significance level, and overall goodness of fit. In the case of the combined model, we tested the hypothesis of common parameters using the likelihood ratio test (equation (4)). Due to the pseudo-panel nature of the SP data, the probability of this component is given by a multiple integral (equation (3)), which allows for a correct treatment of the repeated observations by each individual; in the case of the RP component, the preferred specification turned out to be an MNL.

We also examined the potential correlation among alternatives by estimating NL and ML structures as shown in Figure 1. Specifically, we tested for correlations among private, public, and nonmotorized transport alternatives. However, we could not find reasonable pooled models when we included correlation among alternatives. In particular, it was impossible to obtain meaningful correlations among alternatives in the SP component, and we believe this is a consequence of its experimental design and, in general, poor data collection process. For example, in the choice experiment, individuals had to face eight choice situations but all scenarios presented over five alternatives simultaneously; on hindsight, we believe that this had a nonnegligible impact on the respondent burden (imposing excessive cognitive load onto some respondents), directly damaging the quality of the responses [44].

In the case of the RP component, the difficulties associated with obtaining significant correlations when pooling the models could be due to potential measurement errors in the process used to generate the trip characteristics of the nonchosen alternatives. These issues highlight the importance of using good-quality data, investing time and money in the survey design; they also serve to show the capabilities of using data enrichment techniques when working with data that are not of top quality (which is, unfortunately, the case in many applications, particularly in the Global South).

Equation (5) shows the combined RP-SP model's systematic utility function, and Table 4 presents the definition of the variables finally used in it. The function includes all

Table 3: Summary of main attributes of the individuals included in each dataset.

Attribute	Levels	Description	SP (%)	RP (%)
Age	<22	Under 22 years	11.21	23.13
	22–40	Between 22 and 40 years	62.79	33.27
	40–62	Between 40 and 62 years	25.83	32.39
	>62	Over 62 years	0.15	11.19
Socioeconomic strata (SES)	High	SES 5-6	15.83	2.43
	Medium	SES 3-4	27.97	45.59
	Low	SES 1-2	56.19	51.99
Sex	—	Woman	37.09	47.20
Income	Low	<COP 366,000 (USD 124)	18.18	52.15
	Medium	COP 366,000–COP 2,000,000	71.15	45.43
	High	<COP 2,000,000 (USD 678)	10.66	2.40
Education level	None	The individual did not finish high school	3.10	39.87
	Secondary	The individual finished high school	12.42	35.19
	Tertiary	The individual has higher education studies	87.26	24.93
Occupation: nonoccupied	—	No work or study	8.34	26.62
Household size	—	>3	45.96	51.57
Car availability	—	Yes	52.70	29.49
Motorcycle availability	—	Yes	15.62	13.72
Bicycle availability	—	Yes	33.62	36.58
Trip purpose	—	Work or study related	78.23	29.64
Sample size: individuals			1930	6221
Sample size: observations			15,041	6221

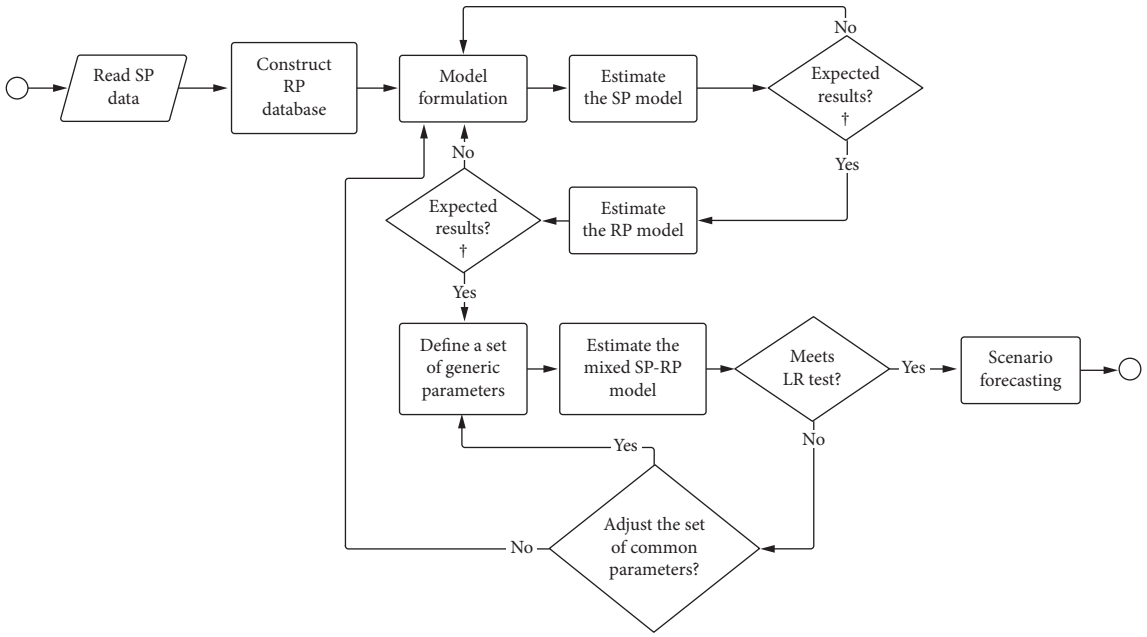

Figure 1: Modelling approach flowchart.

systematic taste variations ([14], page 279) found to be significant in our specification searches, for cost (with occupation and household size), time (with three levels of income, considering the medium level as a base), and the ASC (with sex in the case of public transport). All parameters are common to all alternatives, except sex which was considered specific and was included only in the public transport alternative. The car ASC was taken as reference (i.e., fixed to zero).

$$V_i = \mathrm{ASC}_i + (\beta c + \beta c_{NO} * \mathrm{NO} + \beta c_{HS} * \mathrm{HS}) * C_i + (\beta t + \beta t_{LI} * \mathrm{LI} + \beta t_{HI} * \mathrm{HI}) * T_i + \beta W_i * W * \mathrm{TP}. \quad (5)$$

Table 4: Description of parameters and attributes in the utility function.

Parameter	Name	Associated attribute	Associated attribute description	Expected sign
βc	Cost	C_i	Travel cost (×COP 1000)	–
βc_{NO}	Cost interaction: nonoccupied individual	NO	Dummy that takes the value of 1 if the individual does not work nor study	NE
βc_{HS}	Cost interaction: household size	HS	Dummy that takes the value of 1 if the household size is larger than 3	NE
βt	Time	T_i	Travel time (min)	–
βt_{LI}	Time interaction: low income	LI	Dummy that takes the value of 1 if income is less than COP 366,000 (124 USD)	+
βt_{HI}	Time interaction: high income	HI	Dummy that takes the value of 1 if income is over COP 2,000,000 (678 USD)	–
βW_i	Women	W	Dummy that takes the value of 1 for women	NE
	Public transport	PT	Dummy that takes the value of 1 if the alternative is bus, BRT, train, or metro	NE

NE = no expectations.

In the combined model, we set the ASC to be specific to each alternative and dataset because both environments represented different market conditions and we could not impose correspondence between each market's sample shares.

Figure 2 plots the parameter values from the separate RP and SP models, and shows a line passing through the origin that should approximate the scale ratio between both datasets (recall that we normalized the scale parameter of the RP data to one). Based on this plot and testing with the LR index (equation (4)), we concluded that to be consistent with the partial data enrichment paradigm, the structure of the combined RP-SP model should have specific parameters for cost and for the time interaction with low income in each environment.

Now, even though the approach to define common parameters in the combined model appears clear, some considerations are in order. First, the procedure is visual, so defining what falls outside the elliptic region of acceptance (i.e., away from the line) is subjective. Further, the plotted linear regression curve between both sets of parameters is conditioned by the magnitude of the attributes, which is the primary determinant of the values of the estimated parameters. For instance, the difference between the systematic variations of time could not be visually identified because its values were close to zero, so they might seem to be located inside the elliptical region around the curve.

For this reason, we decided to evaluate the difference between the SP and RP parameters analytically, considering that their ratio should be close to the value of the slope of the fitted line. A large divergence between this ratio and the slope (i.e., the scale) suggests that the parameter is a candidate to be specific in each domain. This procedure complements the graphical evaluation for the definition of common parameters.

Moreover, if this procedure and the graphical evaluation suggest that a pair of parameters should be considered as specific, but one of the estimates from any domain is not significant at the chosen confidence level, the parameter might still be defined as common. In this case, the estimate for the nonsignificant domain is assumed to be equal to its counterpart. However, to be consistent with the data enrichment paradigm, this assumption must be validated by the likelihood ratio test (equation (4)).

7. Results and Discussion

7.1. Modelling Results. Table 5 presents the estimation results for the combined RP-SP model and the separate models estimated using the RP and SP datasets alone. In both cases, the SP component was modeled using an ML structure to consider the pseudo-panel effect. Nonsignificant ASC are not shown in the table (and recall that the ASC for car was used as a reference, and fixed to zero in both environments).

Most parameters of the combined RP-SP model were significant at the 95% confidence level, improving the performance of the separate RP and SP models. The interactions of occupation with cost and of low-income with travel time were nonsignificant in the RP environment. In the first case, the parameter was defined as common; in the second case, it was not possible to assume equal preference with the SP environment since the specification results failed the likelihood ratio test.

The ASC of train and metro were significantly different from zero in both cases, suggesting a higher preference for these new alternatives—*ceteris paribus*—over the car. These results are consistent with the sample shares of Table 3, and may evidence a potential policy bias in the SP responses (i.e., some respondents could have been inclined to choose train or metro because they were in favour of the implementation of these new alternatives); note that the train and metro projects have been in public discussion for decades and have generated great expectations among the Bogotá population.

Note also that as the scale parameter μ turned out to be not significantly different from one, we can deduce that the error variance in both environments is rather similar. Finally, the value of the likelihood ratio test (equation (4)) is 8.79 for five degrees of freedom; this value has to be compared with the critical χ^2 value for a 95% confidence level (11.07). As the LR value is smaller, we cannot reject the

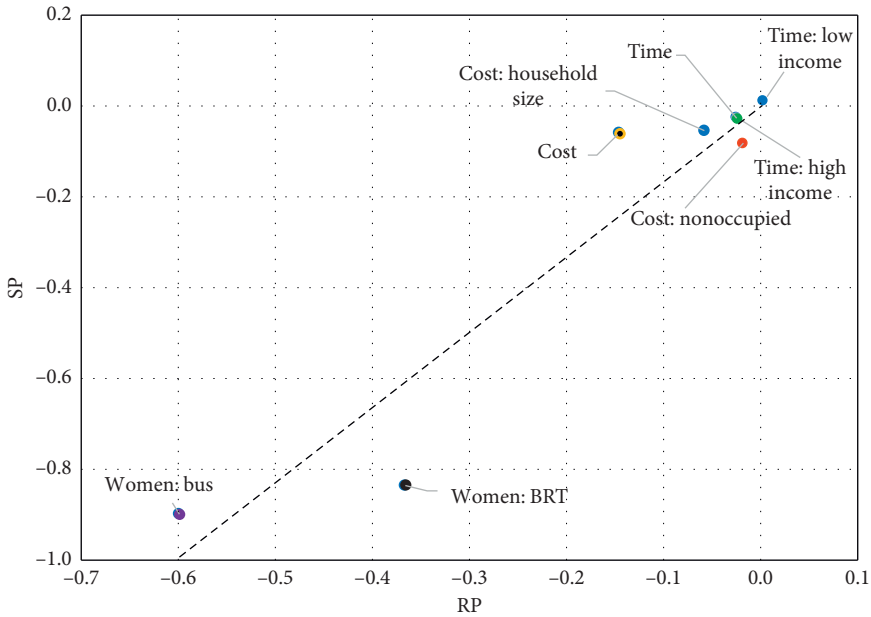

Figure 2: Comparison of potentially common parameters, SP vs. RP.

Table 5: Model estimation results.

Parameter	SP		RP		Mixed RP-SP	
	Estimates	*t*-test	Estimates	*t*-test	Estimates	*t*-test
ASC train (SP)	0.941	5.52	—	—	0.820	4.66
ASC metro (SP)	2.097	12.19	—	—	1.919	8.37
ASC motorcycle (RP)	—	—	−1.312	−12.30	−1.304	−12.26
ASC BRT (RP)	—	—	−0.974	−10.37	−0.925	−10.48
ASC bicycle (RP)	—	—	−2.409	−22.09	−2.424	−22.24
ASC walking (RP)	—	—	0.305	3.37	0.272	3.02
Cost (SP)	−0.060	−3.28	—	—	−0.059	−3.61
Cost (RP)	—	—	−0.145	−8.06	−0.141	−8.97
Cost interaction: nonoccupied	−0.082	−1.94	−0.018	−0.80	−0.035	−1.89
Cost interaction: household size	−0.055	−3.24	−0.058	−3.08	−0.055	−4.48
Time	−0.025	−31.30	−0.025	−12.65	−0.024	−13.11
Time interaction: low income (SP)	0.012	10.31	—	—	0.012	8.59
Time interaction: low income (RP)	—	—	0.002	1.03	0.001	0.55
Time interaction: high income	−0.028	−11.46	−0.023	−2.99	−0.027	−8.87
Women: BRT	−0.834	−5.23	−0.367	−3.26	−0.507	−5.64
Women: bus	−0.899	−5.67	−0.599	−7.71	−0.649	−9.33
Women: train (SP)	−0.994	−6.90	—	—	−0.815	−6.16
Women: metro (SP)	−0.626	−4.70	—	—	−0.468	−4.09
Panel effect std. deviation	1.811	56.96	—	—	1.723	12.42
μ	—	—	—	—	1.051	12.56
Log-likelihood	−17,393.37		−3,960.05		−21,357.81	
Adj. rho^2	0.331		0.180		0.3076	

null hypothesis of the partial data enrichment paradigm (i.e., that the combined RP-SP specification is appropriate) at the 95% confidence level.

7.2. Subjective Values of Time. We estimated subjective values of time (SVT) for the separate and combined models [45, 46], considering all systematic taste variations for cost and time (Table 6). We also computed confidence intervals at the 95% level following Armstrong et al. [47].

In the separate SP model, the point estimates of the SVT range between USD 1.3 and USD 18.1 per hour (COP 3,720 and 53,400, respectively), while in the RP model they only vary between USD 2.1 and USD 6.8 per min (COP 6,240 and COP 20,040, respectively). Apart from the fact that the individuals in the SP dataset have higher income and higher education than those in the RP set (see Table 3), the upper value in the SP case is much higher. It might be related to a potential bias related to the hypothetical nature of the choice scenarios and the aforementioned deficiencies in the design

Table 6: Subjective values of time (USD/hour).

Systematic variations			SP model		RP model		Mixed SP-RP model			
							SP-specific cost parameter		RP-specific cost parameter	
Income	Occupation	Household size	Point est.	Confidence interval*	Point est.	Confidence interval*	Point est.	Confidence interval*	Point est.	Confidence interval*
Low	Occupied	≤3	4.2	2.6–10.1	3.2	2.8–4.7	3.6	2.3–7.5	1.6	1.1–2.2
Low	Occupied	>3	2.2	1.6–3.1	2.3	2.1–3.1	1.9	1.4–2.6	1.1	0.9–1.5
Low	Nonoccupied	≤3	1.7	1.1–4.2	2.9	2.8–4.7	2.3	1.5–4.3	1.3	0.9–1.8
Low	Nonoccupied	>3	1.3	0.9–2.3	2.1	2.1–3.1	1.5	1.1–2.1	1.0	0.7–1.3
Medium	Occupied	≤3	8.4	5.4–20.0	3.5	2.8–4.7	7.6	5.1–16.0	3.4	2.7–4.4
Medium	Occupied	>3	4.3	3.4–6.2	2.5	2.1–3.1	4.1	3.2–5.5	2.4	2.0–2.9
Medium	Nonoccupied	≤3	3.5	2.2–8.8	3.1	2.8–4.7	4.9	3.3–9.2	2.7	2.1–3.6
Medium	Nonoccupied	>3	2.5	1.8–4.6	2.3	2.1–3.1	3.1	2.4–4.5	2.1	1.7–2.6
High	Occupied	≤3	18.1	11.5–43.8	6.8	4.5–9.8	16.3	10.7–34.4	7.2	5.6–9.4
High	Occupied	>3	9.4	7.1–13.7	4.9	3.2–6.8	8.7	6.7–12.0	5.2	4.2–6.3
High	Nonoccupied	≤3	7.6	4.7–19.2	6.0	4.5–9.8	10.4	7.0–19.9	5.8	4.4–7.7
High	Nonoccupied	>3	5.5	3.7–10.9	4.5	3.3–6.9	6.7	5.0–9.7	4.4	3.4–5.5

*95% confidence interval.

of the SP experiment. Previous studies have estimated SVT for Bogotá with an average value of USD 5.2 per hour (COP 15,480) [48], and for public transport and car users between USD 1.4 and USD 5.5 per hour (COP 4,020 to COP 16,140, respectively), depending on the level of crowding inside the vehicle [49]. Although all point estimates fall inside the confidence interval of the SVT for each segment, the intervals for the RP environment have less variability.

7.3. Forecasting with the Models. Using a mixed data model for forecasting purposes requires selecting the parameters that will be used to estimate the probability of choosing each alternative. If all attributes included in the mixed RP/SP model are common, the forecasting process should not involve particular problems in this sense, since the parameters are estimated using data from both environments. However, if this is not the case and some parameters are estimated separately (i.e., are specific) for each dataset, the modeller needs to select which values should be used for forecasting purposes.

Cherchi and Ortúzar [50] discussed the role of ASC when forecasting with combined RP-SP models under the light of reliability and model specification. Their general guidelines suggest that if there are no new alternatives, the RP-ASC should be adjusted to the base year market shares. If there are new alternatives, however, the SP-ASC may be adjusted and used only if there is the certainty that the dataset correctly represents the population market shares. If this is not the case, or if the population market shares are unknown, the estimation results might be used if the model estimates satisfy the microeconomic and behavioural conditions of the modelling framework, such as negative marginal utilities for travel time and cost [50].

Likewise, if specific parameters for each environment are estimated in the joint RP-SP model, to choose the most appropriate values for forecasting purposes, the modeller should prefer those that provide more reasonable and consistent values (i.e., those that represent more appropriately the observed behaviour of users and market shares).

In our case study, the ASC from the SP domain is obviously incapable of reproducing the observed market shares because of the existence of two new alternatives, namely, metro and train (see Table 2). For this reason, we decided to use the ASC from the RP environment for forecasting purposes. However, for predicting scenarios where metro and train are indeed available, there is no other option than to rely on their ASC from the SP environment, which must be scaled by the factor μ [2].

Regarding the specific parameters associated with variables of the individual, mode, or trip, the most reliable estimates should be selected based on signs, significance, marginal rates of substitution, and how sensitive the forecasting results are to each domain-specific parameter. In this sense, willingness to pay and elasticities could be validated with findings from similar studies and their magnitudes contrasted in the light of the socioeconomic characteristics of the study population. Different situations might occur when forecasting with mixed RP-SP models, as follows:

(i) If specific parameters to both domains have adequate signs and significance, the RP parameters should be preferred since they consider real market situations.

(ii) If specific parameters for the same attribute are estimated, but one is not significant or the sign is not consistent with theory, the significant parameter with the appropriate sign should be chosen, regardless of the domain. As in the case of the ASC, if a SP-specific parameter is selected for forecasting, it should be scaled by μ as forecasting always refers to the RP environment [50]. This also applies if a specific parameter is included in only one domain and its marginal utility is consistent with microeconomic and behavioural conditions.

Table 7: Parameter selection for forecasting.

Parameter	Situation	Considerations
Cost	Specific to both domains	The RP-specific parameter was used
Cost interaction: nonoccupied	Common	—
Cost interaction: household size	Common	—
Time	Common	—
Time interaction: low income	Specific to both domains	The SP-specific parameter was used and scaled
Time interaction: high income	Common	—
Women: BRT	Common	—
Women: bus	Common	—
Women: metro	Specific to the SP domain	The parameter was used and scaled
Women: train	Specific to the SP domain	The parameter was used and scaled

Following these considerations, Table 7 presents the selection of parameters used for forecasting in this paper.

With this, the systematic utility functions used to forecast the set of alternatives originally available in the study area are shown in equations (6)–(11), while the functions used to forecast the new alternatives, train and metro, are given by equations (12) and (13).

$$V_{\text{Car}} = (-0.141 - 0.035 * \text{NO} - 0.055 * \text{HS}) * C_i + (-0.024 + 0.012 * 1.051 * LI - 0.027 * \text{HI}) * T_i, \tag{6}$$

$$V_{\text{Motorcycle}} = -1.304 + (-0.141 - 0.035 * \text{NO} - 0.055 * \text{HS}) * C_i + (-0.024 + 0.012 * 1.051 * \text{LI} - 0.027 * HI) * T_i, \tag{7}$$

$$V_{\text{Bus}} = (-0.141 - 0.035 * \text{NO} - 0.055 * \text{HS}) * C_i + (-0.024 + 0.012 * 1.051 * \text{LI} - 0.027 * \text{HI}) * T_i + (-0.649 * W * \text{TP}), \tag{8}$$

$$V_{\text{BRT}} = -0.925 + (-0.141 - 0.035 * \text{NO} - 0.055 * \text{HS}) * C_i + (-0.024 + 0.012 * 1.051 * \text{LI} - 0.027 * \text{HI}) * T_i + (-0.507 * W * \text{TP}), \tag{9}$$

$$V_{\text{Bicycle}} = -2.424 + (-0.024 + 0.012 * 1.051 * \text{LI} - 0.027 * \text{HI}) * T_i, \tag{10}$$

$$V_{\text{Walking}} = 0.272 + (-0.024 + 0.012 * 1.051 * \text{LI} - 0.027 * \text{HI}) * T_i, \tag{11}$$

$$V_{\text{Train}} = 0.820 * 1.051 + (-0.141 - 0.035 * \text{NO} - 0.055 * \text{HS}) * C_i + (-0.024 + 0.012 * 1.051 * \text{LI} - 0.027 * \text{HI}) * T_i + (-0.815 * 1.051 * W * \text{TP}), \tag{12}$$

$$V_{\text{Metro}} = 1.919 * 1.051 + (-0.141 - 0.035 * \text{NO} - 0.055 * \text{HS}) * C_i + (-0.024 + 0.012 * 1.051 * \text{LI} - 0.027 * \text{HI}) * T_i + (-0.468 * 1.051 * W * \text{TP}). \tag{13}$$

We used the cost parameter from the RP domain in forecasting, as we believe it represents more appropriately the subjective valuation of time of users than the SP-specific estimate. We also selected the specific parameter for low income (interacting with time) from the SP environment, as the result from the RP domain was not significant. Note that this parameter was tested as common but the results failed the likelihood ratio test. In contrast, although the parameter for nonoccupied (interacting with cost) was nonsignificant in the RP environment, assuming equal preference between domains was consistent with the data enrichment paradigm, so the parameter was specified as common. As mentioned above, all SP-specific parameters used in forecasting need to be scaled when passed to the RP environment, including the interaction of low income with time, as well as the ASC and gender parameters for the new alternatives, namely, train and metro.

The combined RP-SP model was used to forecast the modal choice impacts of a set of transport projects in line with the public plans for the Bogotá region. In particular, the four following scenarios were simulated using the model, together with data, individual information, and expansion factors from the 2019 Mobility Survey.

Table 8: Market shares by scenario.

Alternative	2019 mobility survey modal split (%)	Scenario 1		Simulations RP/SP			
		RP (%)	SP (%)	Scenario 1 (%)	Scenario 2 (%)	Scenario 3 (%)	Scenario 4 (%)
Car	13.2	13.0	9.8	13.0	12.3	9.5	9.1
Motorcycle	4.2	4.0	5.8	3.7	3.5	3.0	2.8
Bicycle	5.7	6.2	45.4	5.9	5.7	4.7	4.6
Walking	38.0	34.2	-	34.5	34.3	29.8	29.7
BRT	12.6	17.4	19.8	16.5	19.3	13.1	15.3
Bus	26.3	25.2	19.2	26.5	24.9	20.8	19.5
Metro	—	—	—	—	—	12.1	12.0
Train	—	—	—	—	—	7.1	6.9

(i) Scenario 1: business as usual (BAU): there is no introduction of new transport systems, or the construction of regional roads, or any important infrastructure changes. The BRT network remains the same as in 2019.

(ii) Scenario 2: BRT expansion: this scenario includes only new BRT lines (phases 4 and 5,148 new km in total).

(iii) Scenario 3: new rail infrastructure: this scenario includes the first metro line (23.4 km) and two regional train corridors (48 km of the northern line and 39.6 km of the western line). The BRT network remains the same as in 2019.

(iv) Scenario 4: all changes: new public transport infrastructures from scenarios 2 and 3 (BRT lines + rail) are implemented. Also, there is a new regional road network [39].

The simulation of scenarios that include the expansion of the BRT system assumed an improvement of 10% in the average travel time, giving new corridors and connections. The average speed of train and metro was assumed to be 30% higher than that of BRT, following the considerations of the operational design of the railway corridors. The availability of these new alternatives was determined using GIS utilities, depending on the proximity of a station to the origin and destination zones of the trips considered. Note that forecasts were developed considering the simplifying assumption of unchanged road travel times due to modal shifts.

Table 8 compares the current modal split in the city (using the trip data from the 2019 Mobility Survey), with the results of the simulations for all scenarios using the combined SP-RP model. To run the simulations, the cost and travel times of the available alternatives came from the 2019 Mobility Survey, following a similar procedure to that used to construct the RP databank as described earlier.

As seen in Scenario 1, the SP model fails to reproduce the market shares observed in the 2019 Mobility Survey, overestimating the share of the bicycle. Since the SP data did not include walking, the model assigns a significant portion of walking trips to bicycle, since this is the most similar alternative. Most of these trips correspond to short-distance trips for which travel time is similar in all modes, so zero-cost alternatives are preferred.

In contrast, the RP model yields better forecasts, since the conditions of Bogotá's urban and transport system between 2015 and 2019 were similar. There is an overestimation of the BRT market share in the RP data compared to the mobility survey and an underestimation of the trips made on foot. One possible explanation for the above is that our model does not consider some attributes that may be important in mode choice, such as safety and comfort. Another possibility is that some people are willing to walk long distances in Bogotá rather than using BRT, which may be highly attractive in the model but not in reality due to affordability issues [37]. Nevertheless, the mixed RP/SP model provides a better representation of the observed market shares and, more importantly, it allows forecasting the impact of new alternatives in the market. Therefore, building the RP dataset using emerging sources allowed us to enrich the full dataset available, improving the forecasting ability of the choice model to evaluate nonexisting modes in future scenarios.

In Scenario 2, results show an increase in the preference for BRT, which depends on the capacity of the system to reduce travel times. Thus, the expansion of the BRT system implies an increase of about 3% in the market share of this alternative, but its capacity to attract private mode users is limited. In Scenario 3, the new rail infrastructure attracts a significant proportion of travel demand in the city and seems to be more capable to attract car and motorcycle users. This is related to the higher preference *ceteris paribus* of these new modes compared to BRT and bus, as can be inferred from the differences in their ASC. After the implementation of all the new projects (Scenario 4), a decrease in the market shares of car, motorcycle, bus, and walking was expected, as trips transferred to the new alternatives BRT, metro, and train. It is noteworthy, here, that even though train and metro are indeed the preferred alternatives among all user segments, according to the model, as their spatial coverage is limited, they can only capture a portion of the total daily trips made.

In summary, considering that the model will be part of a dynamic Land-Use and Transport model to evaluate urban development scenarios, its forecasting capabilities appear to be robust enough to predict the mode choice of users in future market conditions and different normative contexts.

8. Conclusions

Based on a case study for the city of Bogotá, we have shown that data from emerging sources can be integrated efficiently through the *data enrichment paradigm* to improve the information required to make strategic forecast in scenarios including new alternatives, policies, or changes that affect existing market structures. However, the formulation and estimation of pooled RP and SP models under this paradigm is not a straightforward process, and requires a series of considerations and technical steps that must be strictly followed by modellers to obtain valid results.

In particular, our case study confirmed that in the search for common parameters to both environments, and in the definition of which should remain specific, the graphic approach and likelihood ratio test proposed by Louviere et al. [2] are, indeed, unavoidable steps. Nevertheless, the outcome of the graphic approach is sensitive to the values of the attributes considered and, if followed simply, depends on a subjective visual interpretation. For this reason, we argue that the graphic method needs an analytical evaluation of the proportional differences in the values of the scaled parameters as estimated independently for each domain. We also provide some guidelines on which domain-specific parameters should be used for forecasting when the values for a given policy variable turn out to be not common after the joint estimation and the evaluation of the LR test.

Our case study involved estimating a joint model using two sources of information: an online self-respondent SP survey (unfortunately not carefully designed and, thus, of dubious quality) and a RP dataset based on a large mobility survey complemented by the use of secondary information and GIS utilities, taking advantage of new technological resources like the Google's Distance Matrix API. These tools provide a valuable resource to construct and complement data for discrete choice models, and we encourage researchers to look for these alternatives to enrich discrete choice datasets.

The combined RP-SP model proved to be a more robust tool for travel behaviour analysis and forecasting than the individual RP and SP models. Subjective values of time were also estimated for the independent and combined models, including confidence intervals, finding that the SP component produced noncredible, high values in some cases and wider confidence intervals compared to those obtained from the RP domain (and other related time valuations available for Bogotá). We tested the model for several policy scenarios showing its practical value as a strategic mode choice model to be integrated in a larger Land-Use and Transport Interaction model.

The procedures and implications for forecasting with mixed RP-SP models are still subject to study. In particular, the potential existence of parameters with different values in the RP and SP environments requires the modeller to carefully select which estimates should be included in the forecasting model; this requires considering their reliability, consistency with microeconomic and behavioural theories, and sensitivity in evaluation. In this line, the RP component is generally useful to analyze real conditions of preference by the users, while the SP component usually allows evaluating a wider range of taste heterogeneity and substitution patterns.

However, what to do in the case of new alternatives is still a matter of discussion. Our findings suggest that using the RP parameters for current alternatives provides a good approximation to represent the population market shares. However, for new alternatives, the specific estimates for the SP domain need some caution and could be trusted only if the results are consistent with microeconomic conditions and if the predicted market shares are consistent with rational expectations. At this stage, the expertise and criteria of the modeller, the evaluation of previous experiences, and the availability of similar studies for comparison may play a vital role in the interpretation of results.

Data Availability

The microdata belong to the Bogotá local government. Due to contractual issues, we are not allowed to share the data.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to express their gratitude to the Bogotá Urban Planning Department (https://www.sdp.gov.c#o) for the funding and general support given for the development of this study through agreement 369 of 2018. One of the authors is grateful for the funding by the Instituto Sistemas Complejos de Ingeniería (ISCI) through grant ANID PIA/BASAL AFB180003 and the BRT+ Centre of Excellence (https://www.brt.cl) funded by the VREF.

References

[1] L. G. Willumsen, "Use of big data in transport modelling," 2020, https://www.itf-oecd.org/big-data-transport-modelling.

[2] J. J. Louviere, D. A. Hensher, and J. D. Swait, "Combining sources of preference data," in *Stated Choice Methods*, pp. 227–251, Cambridge University Press, Cambridge, UK, 2000.

[3] E. Cherchi and C. Bhat, "Workshop synthesis: data analytics and fusion in a world of multiple sensing and information capture mechanisms," *Transportation Research Procedia*, vol. 32, pp. 416–420, 2018.

[4] P. Bonnel and M. A. Munizaga, "Transport survey methods - in the era of big data facing new and old challenges," *Transportation Research Procedia*, vol. 32, pp. 1–15, 2018.

[5] N. Rieser-Schüssler, "Capitalising modern data sources for observing and modelling transport behaviour," *Transportation Letters*, vol. 4, no. 2, pp. 115–128, 2012.

[6] E. J. Miller, S. Srikukenthiran, and B. Chung, "Workshop synthesis: household travel surveys in an era of evolving data needs for passenger travel demand," *Transportation Research Procedia*, vol. 32, pp. 374–382, 2018.

[7] G. Harrison, S. M. Grant-Muller, and F. C. Hodgson, "New and emerging data forms in transportation planning and policy: opportunities and challenges for "track and trace"

data," *Transportation Research Part C: Emerging Technologies*, vol. 117, Article ID 102672, 2020.

[8] S. M. Grant-Muller, A. Gal-Tzur, E. Minkov, S. Nocera, T. Kuflik, and I. Shoor, "Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data," *IET Intelligent Transport Systems*, vol. 9, no. 4, pp. 407–417, 2015.

[9] L. Montero, X. Ros-Roca, R. Herranz, and J. Barceló, "Fusing mobile phone data with other data sources to generate input OD matrices for transport models," *Transportation Research Procedia*, vol. 37, pp. 417–424, 2019.

[10] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012.

[11] M. Picornell, T. Ruiz, M. Lenormand, J. J. Ramasco, T. Dubernet, and E. Frías-Martínez, "Exploring the potential of phone call data to characterize the relationship between social network and travel behavior," *Transportation*, vol. 42, no. 4, pp. 647–668, 2015.

[12] G. Romanillos, M. Zaltz Austwick, D. Ettema, and J. De Kruijf, "Big data and cycling," *Transport Reviews*, vol. 36, no. 1, pp. 114–133, 2016.

[13] S. Hankey, W. Zhang, H. T. K. Le, P. Hystad, and P. James, "Predicting bicycling and walking traffic using street view imagery and destination data," *Transportation Research Part D: Transport and Environment*, vol. 90, Article ID 102651, 2021.

[14] J. d. D. Ortúzar and L. G. Willumsen, *Modelling Transport*, Wiley, Chichester, UK, 2011.

[15] J. J. Bates, "Econometric issues in stated preference analysis," *Journal of Transport Economics and Policy*, vol. 22, no. 1, pp. 59–69, 1988.

[16] M. Ben-Akiva, M. Bradley, T. Morikawa et al., "Combining revealed and stated preferences data," *Marketing Letters*, vol. 5, no. 4, pp. 335–349, 1994.

[17] E. Cherchi and J. d. D. Ortúzar, "On the use of mixed RP/SP models in prediction: accounting for systematic and random taste heterogeneity," *Transportation Science*, vol. 45, no. 1, pp. 98–108, 2011.

[18] E. Cherchi and J. d. D. Ortúzar, "Mixed RP/SP models incorporating interaction effects," *Transportation*, vol. 29, no. 4, pp. 371–395, 2002.

[19] D. A. Hensher, J. M. Rose, and W. H. Greene, "Combining RP and SP data: biases in using the nested logit 'trick' - contrasts with flexible mixed logit incorporating panel and scale effects," *Journal of Transport Geography*, vol. 16, no. 2, pp. 126–133, 2008.

[20] E. Cherchi and J. d. D. Ortúzar, "On fitting mode specific constants in the presence of new options in RP/SP models," *Transportation Research Part A: Policy and Practice*, vol. 40, no. 1, pp. 1–18, 2006.

[21] M. A. Bierlaire, *Short Introduction to Pandas*, Biogeme, Ascona, Switzerland, 2020.

[22] S. Hess and D. Palma, "Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application," *Journal of Choice Modelling*, vol. 32, Article ID 100170, 2019.

[23] A. M. Pnevmatikou, M. G. Karlaftis, and K. Kepaptsoglou, "Metro service disruptions: how do people choose to travel?" *Transportation*, vol. 42, no. 6, pp. 933–949, 2015.

[24] X. Yan, J. Levine, and X. Zhao, "Integrating ridesourcing services with public transit: an evaluation of traveler responses combining revealed and stated preference data," *Transportation Research Part C: Emerging Technologies*, vol. 105, pp. 683–696, 2019.

[25] W. Li and M. Kamargianni, "Providing quantified evidence to policy makers for promoting bike-sharing in heavily air-polluted cities: a mode choice model and policy simulation for Taiyuan-China," *Transportation Research Part A: Policy and Practice*, vol. 111, pp. 277–291, 2018.

[26] Z. Rashedi, M. Mahmoud, S. Hasnine, and K. N. Habib, "On the factors affecting the choice of regional transit for commuting in greater toronto and Hamilton area: application of an advanced RP-SP choice model," *Transportation Research Part A: Policy and Practice*, vol. 105, pp. 1–13, 2017.

[27] J. d. D. Ortúzar and C. Simonetti, "Modelling the demand for medium distance air travel with the mixed data estimation method," *Journal of Air Transport Management*, vol. 14, no. 6, pp. 297–303, 2008.

[28] P. Lizana, J. d. D. Ortúzar, J. Arellana, and L. I. Rizzi, "Forecasting with a joint mode/time-of-day choice model based on combined RP and SC data," *Transportation Research Part A: Policy and Practice*, vol. 150, pp. 302–316, 2021.

[29] R. Paleti, P. S. Vovsha, D. Givon, and Y. Birotker, "Joint modeling of trip mode and departure time choices using revealed and stated preference data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2429, no. 1, pp. 67–78, 2014.

[30] H. Resano-Ezcaray, A. I. Sanjuán-López, and L. M. Albisu-Aguado, "Combining stated and revealed preferences on typical food products: the case of dry-cured ham in Spain," *Journal of Agricultural Economics*, vol. 61, no. 3, pp. 480–498, 2010.

[31] J. de Dios Ortuzar and A. Iacobelli, "Mixed modelling of interurban trips by coach and train," *Transportation Research Part A: Policy and Practice*, vol. 32, no. 5, pp. 345–357, 1998.

[32] D. L. McFadden, "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics*, pp. 105–142, Academic Press, New York, NY, USA, 1973.

[33] K. E. Train, *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, UK, 2009.

[34] L. A. Guzman, D. Oviedo, and C. Rivera, "Assessing equity in transport accessibility to work and study: the Bogotá region," *Journal of Transport Geography*, vol. 58, pp. 236–246, 2017.

[35] L. A. Guzman, D. Oviedo, and J. P. Bocarejo, "City profile: the Bogotá metropolitan area that never was," *Cities*, vol. 60, pp. 202–215, 2017.

[36] V. Cantillo-García, L. A. Guzman, and J. Arellana, "Socio-economic strata as proxy variable for household income in transportation research. Evaluation for Bogotá, Medellín, Cali and Barranquilla," *DYNA*, vol. 86, no. 211, pp. 258–267, 2019.

[37] L. A. Guzman and D. Oviedo, "Accessibility, affordability and equity: assessing "pro-poor" public transport subsidies in Bogotá," *Transport Policy*, vol. 68, pp. 37–51, 2018.

[38] L. Guzman, D. Oviedo, and R. Cardona, "Accessibility changes: analysis of the integrated public transport system of Bogotá," *Sustainability*, vol. 10, no. 11, p. 3958, 2018.

[39] L. A. Guzman, F. Escobar, J. Peña, and R. Cardona, "A cellular automata-based land-use model as an integrated spatial decision support system for urban planning in developing cities: the case of the Bogotá region," *Land Use Policy*, vol. 92, Article ID 104445, 2020.

[40] L. A. Guzman, J. Arellana, and V. Alvarez, "Confronting congestion in urban areas: developing sustainable mobility plans for public and private organizations in Bogotá,"

Transportation Research Part A: Policy and Practice, vol. 134, pp. 321–335, 2020.

[41] L. A. Guzman, "A strategic and dynamic land-use transport interaction model for Bogotá and its region," *Transportmetrica B: Transport Dynamics*, vol. 7, no. 1, pp. 707–725, 2019.

[42] A. Daly and J. d. D. Ortúzar, "Forecasting and data aggregation: theory and practice," *Traffic Engineering and Control*, vol. 31, pp. 632–643, 1990.

[43] M. D. Ribeiro, A. M. Larrañaga, J. Arellana, and H. B. B. Cybis, "Influence of GPS and self-reported data in travel demand models," *Procedia - Social and Behavioral Sciences*, vol. 162, pp. 467–476, 2014.

[44] S. Caussade, J. d. D. Ortúzar, L. I. Rizzi, and D. A. Hensher, "Assessing the influence of design dimensions on stated choice experiment estimates," *Transportation Research Part B: Methodological*, vol. 39, no. 7, pp. 621–640, 2005.

[45] M. J. I. Gaudry, S. R. Jara-Diaz, and J. d. D. Ortuzar, "Value of time sensitivity to model specification," *Transportation Research Part B: Methodological*, vol. 23, no. 2, pp. 151–158, 1989.

[46] M. Sillano and J. de Dios Ortúzar, "Willingness-to-pay estimation with mixed logit models: some new evidence," *Environment and Planning A: Economy and Space*, vol. 37, no. 3, pp. 525–550, 2005.

[47] P. Armstrong, R. Garrido, and J. D. D. Ortúzar, "Confidence intervals to bound the value of time," *Transportation Research Part E: Logistics and Transportation Review*, vol. 37, no. 2–3, pp. 143–161, 2001.

[48] M. Gutierrez-Torres and V. Cantillo-Maza, "Classic and Bayesian estimation of subjective value of time," *DYNA*, vol. 81, no. 187, pp. 158–166, 2014.

[49] M. Batarce, J. C. Munoz, J. d. D. Ortuzar, S. Raveau, C. Mojica, and R. A. Rios, *Evaluation of Passenger Comfort in Bus Rapid Transit Systems*, Inter-American Development Bank, Washington, DC, USA, 2015.

[50] E. Cherchi and J. d. D. Ortúzar, "Use of mixed revealed-preference and stated-preference models with nonlinear effects in forecasting," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1977, no. 1, pp. 27–34, 2006.