Big data analytics for modelling consumer preferences and satisfaction in public transportation

Yulia Dzhabarova[1] Aygun Erturk[2] and Stanimir Kabaivanov[3][0000-0002-8686-8112]

[1] Plovdiv University "Paisii Hilendarski", Plovdiv Bulgaria
[2] Plovdiv University "Paisii Hilendarski", Plovdiv, Bulgaria
[3] Plovdiv University "Paisii Hilendarski", Plovdiv, Bulgaria
stanimir.kabaivanov@uni-plovdiv.bg

Abstract. In this paper we develop a model to estimate and analyze consumer preferences and satisfaction from public transportation services. Unlike many other studies in this area, our approach is based on use of big data from multiple sources and allows to achieve continuous and precise estimation of consumer behavior. These results can be used then to adjust parameters of the transportation plans, schedules and asset allocation. We build the model with available data from INNOAIR project in Sofia.

Keywords: big data analytics, consumer behavior, customer satisfaction.

1 Introduction and methodology

1.1 Literature review

The European Standard classified quality criteria of service in public passenger transport into the categories of: availability, accessibility, information, journey time, customer care, comfort, safety and environmental impact ([1]).

There are a large set of service quality attributes. Examples can be given as network design, service supply and reliability, comfort, fare, information, safety, relationship with personnel, customer preservation, environmental protection, quality of exterior, but there is the necessity to quantify the importance of each one as each of them has a different weight ([2]). Also, demographic characteristics play a direct role on perceived service quality ([3]).

There are many service quality measure techniques and the evaluation can be done by customers satisfaction, expectation of the customer for statistical analysis, indexes for measuring or linear and non-linear models. A popular one is SERVQUAL a 22-item instrument for assessing customer perceptions in service and relating organizations, where tangibles, reliability, responsiveness, assurance, and empathy are the dimensions that has an impact on consumers evaluate service quality ([4]). The alternative of SERVQUAL as SERVPERF which is a performance-based measure ([5]).

Customer Satisfaction Index points out to customization and customer expectations and pays attention to the fact of quality than the price for customer satisfaction ([6]).

There are also linear and non-linear models as structural equation model (SEM) which has become one of the most applicable methods in public transport ([7]). Generally service quality measures for public transport are dependent on the perception and expectations of the customer.

1.2 Analytical framework

Consumer behavior is a very broad topic and can be addressed from different points of view. We adhere to the definition that consumer behavior is the study of how people decide on taking actions that satisfy their needs, wishes and desires.

It should be noted that this definition refers to a broader set of options than just making purchases or spending money. For the purpose of INNOAIR[1] project, it is justified to address also actions that do not result in immediate economic spending, due to the fact that public transportation plays a very special role in providing services and support for different social groups.

In accordance with the definition of passenger behavior, used in previous reports, that *"A passenger behavior is the way that passenger think, feel, reason, make judgement and select different products and services, directly related to their travel."* we only select consumer behavior aspects related to public transportation:

- Public transportation and the potential of establishing a green on-demand transportation system.
- Congestion and load on the city road infrastructure.
- Air pollution and emissions in different urban areas.
- Attitude toward public transportation.

All these items have direct impact on the quality of service (QoS) of the public transportation**.** Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

The aim of this study is to distinguish passenger segments with similar behavior, and to identify the public transport they tend to use, and thus to characterize their preference patterns.

Based on previous studies on understanding passenger patterns in different countries and on different occasions, the most important characteristics that distinguish passenger behavior have been determined. The passenger profile is a significant determinant. which lays in the overall behavior, e.g. consumer needs, motives/preferences, attitudes and expectations. After interlinking these variables, we will be able to obtain in-depth understanding of passenger insights and thus to design an empirical model. The study will be focused on the average traveler in the piloting residential districts of Manastirski Livadi and Buxton in Sofia. Different scenarios will be drawn, and appropriate recommendations on adapting "innovative green mobility solutions" to traveler's behavior will be elaborated.

[1] https://innoair-sofia.eu/en/

In order to capture different points of view on consumer behavior, we have conducted a brainstorming session with the aim to build a mind map of the targeted issue and potential ways to solve it. To make sure that sufficient diversity in expertise involved is indeed present, Table 1 provides a summary of special areas of knowledge present in the discussion.

Table 1. Experts involved in the mind mapping and brainstorming session.

Experts	Details and remarks
Marketing	Inputs on standard methods used for marketing study of consumer behavior.
Economics	Inputs on economic and financial aspects of consumer behavior and metrics available to assess them.
Urban development	Inputs on contemporary aspects of urban development as well as on metrics available to clarify existing challenges.
Software development	Inputs on technical solutions and required data to estimate and track down selected metrics.
Public transportation services	Inputs on viability of suggested metrics and mapping to already collected information for public transportation system

Three major areas of interest have been identified with regard to consumer behavior, in the context of INNOAIR project goals:

- Consumer characteristics and profiling.

Consumer characteristics and segmentation goal follows well-defined and commonly used methods and inputs to analyze passenger choices and spending. In addition to traditional tools and features used to study customer decisions, we suggest applying state-of-the-art machine learning methods to improve segmentation results and handle special cases (e.g. outliers).

- Technical aspects.

Technical aspects of the output aim at finding the most relevant and precise sources of data. While it is often sufficient to conduct representative surveys of consumer preferences, such an approach is limited in terms of providing adequate real-time information. For the purpose of accuracy, it is better to combine different sources of information and merge survey results with objective and real-time feeds like fleet management data, anonymized mobile network load and coverage information and ticket sales.

- Implementation details.

Implementation details refers to the use of appropriate technologies and employing open source packages in order to achieve sustainable and economically sound effect.

This part follows from the previous two and despite playing supportive role, it is important to make sure that implementation can fulfill all required actions and process inputs in a timely and accurate manner. Apart from concerns regarding data analysis, it is of great importance to ensure that different sources can be used together and can be addressed in a uniform way. Due to the fact that inputs often come from separate organizations, with different level of technical readiness and degree of process automation, being able to integrate all existing solutions with ease is crucial.

2 Data inputs and theoretical background

2.1 Demographic characteristics

For appropriate consumer profiling, and based on analysis of previous research on public transport in different countries, ([8], [9], [10]) the following key demographic features have been selected: age structure, household size, occupation, personal income, education, vehicle availability.

A**ge structure** of the passenger profile represents types of activities according to the life cycle stages. According to the set-up purposes the passengers are grouped as follows: 15-24, 25-29, 30-40, 41-50, 51-64, and 64+. By 2021 the population of Sofia is 1 307 439 people, and most of them are between 20-64 years old (825 370 people) ([11]), the population projection for Sofia for 2025 is to reach 1 350 054 people [12], while the projected age dependency ratio for Sofia for 2025 is 48.84 while for Bulgaria it is 59.26 ([13]). Age as a factor provides the opportunity to divide life-cycle stages into: education, work, and retirement, and to describe the contrast between the financial opportunities, passenger preferences, and to reveal more precisely the purposes and frequencies of traveling. This division also specifies the travel purpose, the presence of an additional passenger, car ownership, the ability to use alternative modes of travel. Generation differences have also a significant role on passenger preferences. There is a new trend as slow mobility especially among ageing population and interest in modern, healthy, technological and green mobility among the younger ones.

Household size is used as main indicator for the ownership and dependency on personal cars, as it shows the number of working members of the household and appears to explain the choice of different types of transport. The average household size is decreasing and in 2020 both for the European Union and Bulgaria was 2.3 people per household ([14]). The tendency for the number of cars per household is increasing, but the last findings display a decrease in the stock of vehicles in Bulgaria from 3 667 787 in 2015 to 3 339 725 in 2019 [15]. Meanwhile, in Sofia the number of cars is increasing, there are 663 cars per 1000 inhabitants in 2020 ([16]). For the purposes of the research, the household size will be divided into the next groups: a single person, two-person household, three-person household, four-person household, and more, in order to describe the socio-economic groups and their transport preferences.

Occupation of the passenger modifies the time preferences, type of transport, frequency, and also the ticket type as it describes the regular and habitual activities. Occupation options include the next categories: full or part-time employee (managers; professional technicians and associate professionals; clerical support workers and

armed forces occupations; agricultural, craft and related trades, service and sales workers, machine operators, and assemblers) unemployed person, pupil, student, housewife, and retiree. The availability of public transport and the time-lasting of traveling are among the key factors for accepting a job offer. The passenger occupation profile is an indicator and prerequisite to schedule the public transport route **(** [17]**).**

Personal income is used to shape the social demographic structure and is an evidence for special fare usage and car ownership ratio. The structuring of the attitude towards public transport and accordingly passenger preferences, depends on the income range. The passenger belonging to the low, middle, and high-income group will have explanatory value for his habits and attitudes. As a common belief, bus riders rather belong to the lower-income group than the passengers of metropolitan. In Sofia, there is a significant difference between the percentage of metropolitan use between 2011 (9.7 %) and 2020 (24.3 %) ([16]). The passenger income prevails the preference between owning a private car, using public transport or using other free transport types as walking or cycling, and moreover, it motivates the passenger willingness to pay for better quality.

Education level is another main demographic shaping the consumer profile related to the average habitant level. The groups are outlined as: Bachelor level or higher, high school, less than high school education, secondary school. Generally, metropolitan riders are considered as more educated than bus riders. Furthermore, the general aim of the public transport usage is to make it preferable mainly for students and pupils. A distinguishing feature of pupils is that they travel mainly to school and typically travel at a shorter distance. Education level has to be considered for the recommendations seeking to change the general attitudes.

Vehicle availability in the household also effects the passenger preferences of public transport. The increasing level of car ownership is in the basis of the infrastructure and traffic problems. Dependency on own car is influenced by the changing life style and the generation differences.

2.2 Behavioral characteristics

Although previous research results estimate trip purpose, the age and duration as major factors, other ones, including psychological factors, underline flexibility, safety, comfort and environmental concerns as central factors ([18]). In this research passenger profiling implies particular transportation features as frequency and reasons for public transport use, types of sources of information, access method ([8], [19])

- Important aspect of the frequency of using public transport is to make contrast between routine commute behaviors and random on-occasion transport demand.
- The reasons for using public transport as explanatory for behavioral motives to design a need or preference base model can be found in: saving money and time, avoiding traffic, environmental concerns, convenience.
- Access mode to reach the bus stop indicates if the riders walk, drive, bike or use another transit.
- The travel routine of passengers is described by trip the duration and trip road.

- Information sources are important to predict the pre-travel planning behavior of passengers. Real Time Passenger Information is one of the facilitating systems, pre-travel information has also positive psychological effect on the passengers ([20]).

2.3 Social class and lifestyle

Transport practices are ranked among the most habitual behaviors and the older generations are supposed to adhere to car ownership and valuing the old pattern of "car pride" as using car as a symbol of status and prestige and the drivers are not reluctant to prefer public transport ([21]). Yet in EU the belief is that youngest generations will be more interested in technological gadgets and social networks rather than owning car ([9]). Ultimately there is tendency of using the alternative technological and green mobility in urban areas with regard to healthier and sporty lifestyles and increasing health and environmental awareness. The outdated pattern that the public transport is used by lower income group is challenged by generation lifestyle change and use of metro.

Although the existing interdependence between the quality of public transport and income level of passenger, the image campaign, adequate guidance, and information systems play a significant role in changing this pattern ([22]). Our previous passenger behavior analysis report put the major goals to identify needs for public transportation as adaptive schedule and bus lanes, fixed lanes optimization and bus stop location optimization; traffic lights and special rules for public transportation vehicles. This improvement in public transport quality would attract the higher income population with definite requirements.

2.4 Behavioral determinants (passenger insights)

For the purpose of the analysis we use several variables that determine passenger behavior, such as: consumer needs, motives and preferences, attitudes and expectations. Based on the consumer insight approach, and according to the specifics of the distinguished segments we aim to reveal the personal drivers for a concrete behavior.

Needs and motives

In the last years the lifestyle patterns have substantially changed, generating diversified travel needs. People become more concerned about the implications from transport congestion and pollution. The now-a-day consumers tend to be more sensitive to their personal and family health-being, social welfare and ecological issues. Thus, the implementation of holistic behavioral approaches, satisfying new-consumer demands, and gaining social, economic and ecological benefits is getting more and more insistent.

According to the existing requisites the needs can be evaluated as utilitarian or hedonic, Under the utilitarian need lays the desire to achieve practical (functional) ben-

efits, e.g. to sustain innovative mobility solutions, to contribute to the city carbon reduction, to save money, to obtain a quick access to a definite place/point, to avoid traffic conjunction, to care about personal or family's health, etc.; The hedonic needs refer to the emotional striving, e.g. to enjoy the travel, to share the experience of travel with other passengers, to be eco-friendly, to demonstrate a typical lifestyle or social class affiliation, to have personal sentiments to the district/city as a resident or as a place of birth, etc.

In the usual case the consumer need contains both utilitarian and hedonic features, and it can be satisfied in different ways, as the personal choice depends on a specific set of experiences and personal characteristics (demographic, social or cultural, for example the cultural values he has been raised). Hence, the traveler will perform his choice, driven by his want to satisfy a certain need, to attain a concrete goal. To be more precise, we also have to consider the concrete reason for a travel: how it performed - on a regular basis as traveling from home to work/school, or on a specific occasion – to visit friends, to sport, to go for a sight-seeing or amusement, etc. The traveler's occasion will influence the frequency of using different modes of transport – public, private, alternative, etc., and a combination between them.

In order to enhance the public transport use, we need to know the motives that drive passenger behavior, on one hand, and the barriers that prevent him. Table 2 provides a summary of some important motivating and de-motivating factors that influence the decision whether and how to make use of available public transportation system. We have paid more attention to the barriers as most of them are easier to quantify and assess.

Table 2. Motivations and barriers to public transport use (Source: [23])

Motivations	**Barriers**
Better service	Not having alternative to car
Be certain that the timetables are performed	Lack of direct transport
	Lack of availability of buses
Direct transport from home to work	Long travel time
More information available and easy to understand	Buses' unreliability
	Do not known what to expect
Save money	Need for multiple journeys
Not having a parking space	Poor information
More comfort and air-conditioning on vehicles	Not frequent enough
	Bus stop too far
Contribute to a better environment	Buses are smelly and crowded
	Feeling of personal insecurity
	Having to use more than one transport
	Bad waiting conditions
	Negative feeling towards public transport
	Habit of driving

Preferences

The choice of transport is influenced by several factors, such as individual characteristics and lifestyle, the type of journey, the perceived service performance of each transport mode and situational variables ([23]).

Nowadays people perform a private car dependence, not only as a mode of travel, but also to express their social status and hedonic desires, such as feelings of sensation, power, freedom, status and superiority ([24]). In order to make a shift from private cars to other travel modes, they need alternatives, to be convinced in their benefits and corresponding to their personal values and lifestyle. Moreover, innovations in improving the quality of public transport will satisfy the specific traveler' needs and sustain a modern urban transportation system.

Regarding the type of mode, a passenger could choose among 4 options of public transport in Sofia: bus, trolleybus, tram and metro lines. Additional options as taxi, shuttle and electric scooters could be used as well. In order to perform their choice, passengers evaluate the characteristics of the different modes, make their preferences and actually their choices.

Some of the most significant transport characteristics that determine passenger preferences and choice, and influence his satisfaction are the following:

- Convenience;
- Safety;
- Connectivity;
- Reliability;
- Service quality;
- Fastness;
- Information-equipped.

All these choice determinants can be integrated by implementing modern transport systems with improved traffic management. Specific solutions as introduction optimization routes and bus schedules, new eco-bus lines, creation of bus only lanes, innovations and incentives for passengers, digitalization of passenger tickets, etc. may sustain the passenger modern needs and desires.

As long as there are many alternatives, bonded with appropriate information, a passenger will be more facilitated and convinced in his decision. In order to boost this process targeted information campaigns in social media could promote the benefits of sustainable travel modes as eco-friendly, as easily accessible and reliable.

Attitudes

Beliefs and attitudes lay in the entire process of consumer decision making. They are determined by (1) the existing knowledge from the obtained information and gained experience, (2) the emotions evoked by the explored situation, and (3) the beliefs related to the concrete object. The mutual performance of the three variables, as: cognition, affect and behavior, determine the consumer/passenger attitude, by which we

could predict the forthcoming reaction. In the ABC model Michael Solomon ([25]) explains the relative impact of the three components as hierarchy of effects, where each hierarchy specifies that a fixed sequence of steps occurs on the route to attitude, e. g. an attitude based on cognitive information process, an attitude based on behavioral learning process. and an attitude based on hedonic consumption. In other words, passenger behavior could be predicted when the attitudes are evaluated as an output of the hierarchical process of the effects. In this way two scenarios could be performed: (1) beliefs-affect-behavior, or (2) beliefs-behavior-affect. In the first case consumer is motivated to search for a lot of information and carefully evaluates the offered alternatives, and thus comes to a thoughtful decision. In the second case a passenger may not be particularly informed/knowledgeable about the different transport options and public modes, then he may have an emotional response. In this hierarchy a person does not initially have a strong preference for one over another alternative. Even more, when a passenger has "bonds" with a certain transport mode over time, he is not easily persuaded to experiment with other options. Additionally, we have to consider the passenger motives (rational and emotional), the existing knowledge and previous experience to model his attitudes towards the different travel options (public and private) and the different public travel modes. In order to make consumers more involved and conscious about their decision making, the marketing stimuli, affecting the whole process, have to be addressed properly and adequately. Taking in mind that attitudes are time-lasting, a special focus should be given on the communication set of tools, influencing cognitional learning rather than behavioral learning.

Considering the **Expectancy-Value Attitude** model (multi-attribute attitude model), the passenger attitude towards the different modes of public transport might be determined by scaling two factors: salient beliefs about the values possessed and their evaluation ($A_{ob_i} = \sum_{i=1..n} be_i$). The personal evaluation might regard a certain transport mode in general, as it consists of particular evaluation of the demanded characteristics.

Expectations and satisfaction

Usually, passenger satisfaction is conceptualized as a function of the gap between expected and experienced service delivered. Travel satisfaction in transportation research has been predominantly measured as a function of objective or subjective attribute levels ([26]).

The discrepancy between expected and experienced service, as: waiting time, in-vehicle time, perceived service quality, as well as the passenger personality, socio-demographic and behavioral characteristics might have a significant influence on the level of satisfaction. Loyalty to public transport, price affordability, seat availability, in-vehicle environment, enjoyment mood, etc. lead to higher ratings of trip stage satisfaction. Some studies of trip satisfaction found that gender has no significant effect on satisfaction with public transport services ([27], [28]), others ([26]) in reverse state that satisfaction ratings of men are higher than satisfaction ratings of women.

2.5 Technical aspects (external input)

There are multiple data sources that can be used to analyze passenger behavior and use of public transportation system. But not all of them are equally important and useful. In order to find the balance between usability, accessibility and information value, we have separated different options with regard to:

- type of data inputs (numerical, categorical and nominal);

Giving preference to numerical inputs allows us to apply wide variety of analytical tools and provide more accurate and timely forecasts, conclusions and recommendations. While some sources are by virtue either categorical or nominal, we have tried in our report to focus on numeric sources and extract as much useful details as possible.

- frequency of new data that is made available;

Making strategic decisions takes time and preparations. Yet there are some daily changes that have to be made and that require to have information quickly, or its value will simply fade away before its even taken into consideration. In order to support both strategic and tactical decisions, our consumer behavior analysis approach relies on data inputs that arrive at different frequency – from real-time data of movements and fleet positions, to periodic surveys that take several weeks to complete and process.

- accessibility and control of municipal authorities on the quality of methodology of data collection.

Accessibility of data is a twofold issue – it depends on both legal and practical factors. For example, tracking movement of people is a very valuable input, but on the other hand it violates their privacy and intrinsic human rights. To eliminate issues with GDPR and access to sensitive information, in our study we rely on anonymized and pre-processed information that prevents identifying any individual based on the collected data.

With regard to practical concerns, we choose inputs that cannot be manipulated, have negligible margin for error (mostly due to the fact that they are automatically generated and do not require human intervention) and that can be reproduced and verified. The last two characteristics are of particular important in order to make sure that obtained results are valid and sustainable.

Table 3 contains a detailed list of inputs from the following categories:

- mobile network data;

Mobile network data is not freely available and requires close cooperation with mobile service providers. Use of anonymized or preprocessed information on the other hand makes it much easier to cope with legal restrictions and avoid violation of individual rights. To minimize the investment of time and resources on mobile providers' side, we plan to use information on density of users and summary of movements in specific areas, that are relevant to INNOAIR scope.

- ticketing system data;

Ticketing system data is already available and used to estimate revenues and cash flows. In our case it's more important to figure out relative weight of different products (ticket and prepaid travel cards), rather than the absolute amounts. With the introduction of electronic ticketing and check-in system, required inputs are readily available and

can be provided with minimal efforts. To analyze the consumer behavior, it is required to have only timing and general information, thus avoiding potential conflicts with GPDR and use of personal data.

- mobile application statistics;

Mobile application inputs are focused on INNOAIR experimental services and green transportation. Therefore, its relevance is extremely high. Mobile application offers a direct channel of communication with passengers that rely on the service, can rate it and give suggestions on how to improve the quality of offered services. Our goal is to use this data with gradually putting more weight on it, as the number of application users increases.

- traffic and transportation control data.

Traffic and transportation control data is crucial to mapping available resources to demanded services and improving customer satisfaction. This information serves analytical purposes and supports decision making process. In our model, traffic and transportation data is first used to assess current situation and compare it with user expectations. As gradually new services start to gain popularity, estimate that its role will be primarily as a support tool when changes have to be made to meet new requirements and consumer needs.

Table 3. Data inputs and sources.

Data input	Description
	Mobile network inputs
Anonymized position	Used for heat maps and summarized tracking of movement in different hour zones.
Speed of movement	Used for rough estimates and classification of means of transportation
Time of appearance	Entry and exit time, calculated from available movement data as proxy of transportation use and load estimates.
	Ticketing system inputs
Transportation cards	Used to track down established route and sustainable interest in particular lines or areas of transportation. Inputs regarding sales of transportation cards are important not just for economic planning and financial forecasting, but also for separating long-term interest in particular routes.
Tickets sold/checked-in	Tickets sold and/or checked-in can be used to track down in real-time the load of transportation network. It can be used for adaptive scheduling and planning of resource allocation. This information is also very interesting with regard to traffic control and pollution monitoring.
	Mobile application statistics
Number of requests	In addition to number of requests (which is used to heat map the activity and user needs at given time), the distribution of these requests is very useful in resource optimization.
Frequency of use	Used as a proxy for customer satisfaction and segmentation of frequent/devoted users of the services offered.

Feedback and quality study	While this is a voluntary input (not obligatory to rate the service and application) and it provides an assessment of the complete service, quality feedback is a very strong (though irregular) input.
	Traffic and transportation control center
Schedules	Fixed schedules are used to compare demand for transportation services with supply. They are also the foundation of providing more flexible and efficient allocation of assets and forming dynamic schedules.
Delays	Delays are very important in reducing consumer satisfaction (as pointed out in D5.3.1) but also as indicator of problems and deficiencies in planning and implementation of transport network.
GPS live positioning	Live positioning information on vehicles (busses and other means of transportation) is vital for providing quality service and also map consumer behavior to availability of offered services.

Ingestion of different sources and analyzing the data can be split into several steps, as shown on Figure 1. Integration of multiple data sources (which normally deliver inputs with disparate frequencies) is handled first. By abstracting this initial (pre)processing, we are able not only to support series with different frequencies, but also different communication protocols. Therefore, this step is crucial in putting together legacy systems and data producers managed by separate legal entities. Having a separate data ingest layer also helps in simulating some inputs that are not yet available or developed.

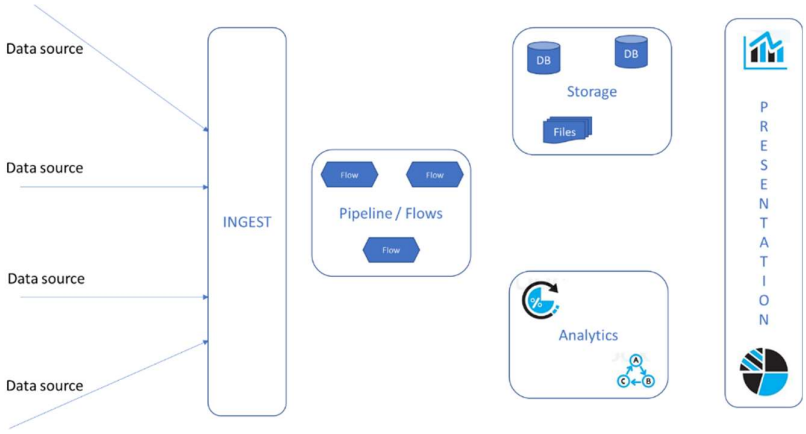

Fig. 1. Data processing and analytics.

Pipeline and flows are central to our way of consumer behavior analysis and monitoring. Representing a sequence of tasks (pre-processing, analytical and data persistence ones), they are capable of wrapping up different algorithms and methods. By

sticking to the idea of flow, that takes predefined inputs, processes them (which could involve multiple steps of different complexity) and stores and/or presents the results, we are able to create building blocks for consumer behavior modelling and link these blocks in very creative ways. This approach can provide the much-needed flexibility and improve sustainability of suggested solutions, as we are always able to modify or adapt some part of them, without rebuilding the complete system. In addition, the ability to immediately present the results is very important for maintaining real-time notifications and "live" view on consumer behavior.

The last part represents the graphical front-end of the system. It is responsible for providing consumer behavior analysis in a readable and convenient way. To highlight the dynamic nature of the built system we have implemented this component as user-controlled and user-configurable dashboard solution. This provides for interactive output and customization based on specific end-user needs.

2.6 Implementation details

Implementation of consumer behavior relies on software packages that fulfill the following generic requirements:

- open source and no licensing spending required for use;
- support by major cloud providers and easy scalability;
- integration with common scripting languages and in particular with R and Python;
- support for common authentication methods and state-of-the-art security mechanisms;
- support for machine learning and AI algorithms.

Figure 2 provides an overview of common packages used to implement different stages of our passenger behavior analysis. There are multiple variants and alternatives to selected software, but we have decided to bet on proven and easy-to-use options.

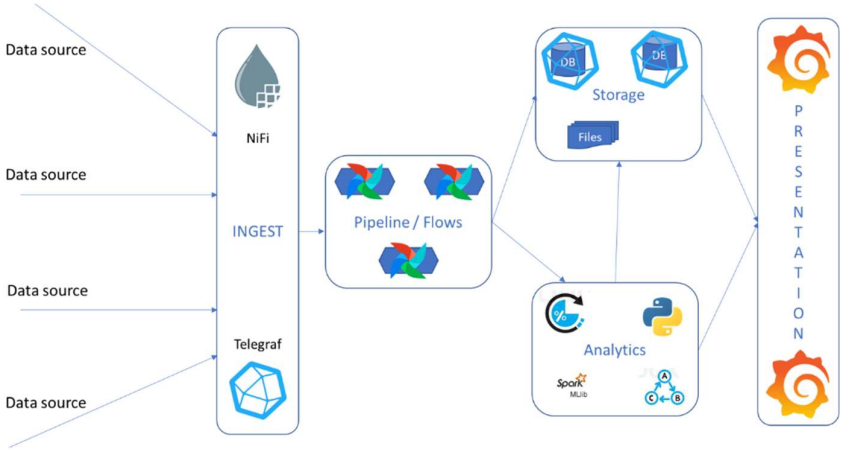

Fig. 2. Software packages used to support analysis.

Table 4 contains a list of selected packages, followed by a short description. Only specialized packages are listed, where common applications (like spreadsheets or survey automation packages) are not explicitly described.

Table 4. Software packages selected for implementation.

Selected solution	Description
	Ingestion stage
Apache NiFi ([29])	Data routing and transformation in NiFi is especially valuable for integrating legacy systems and various inputs.
InfluxDB Telegraf ([30])	Used to collect data and metrics from different devices, which do not need special handling and preprocessing.
	Pipeline / Flows
Apache AirFlow ([31])	Used to orchestrate different tasks and worker threads of the analysis.
	Storage
InfluxDB ([30])	Used to store time series data and do simple calculations over it.
File system	Used to store configuration and binary objects.
	Analytics
Apache Spark MLib ([32])	Used to provide reference machine learning algorithms.
Custom Python scripts	Used to customize and automate tasks, that cannot be easily handled by other software packages.
	Presentation
Grafana ([33])	Used to create interactive dashboards that present study findings as well as any real-time outputs.

2.7 Preliminary results

Due to the fact that collecting sufficient number of observations requires a significant amount of time, our initial tests were conducted with a lot of simulated inputs, aiming mostly at confirming that implementation can meet flexibility and sustainability expectations. Our preliminary results of the system agility can be summarized as follows:

- Thanks to the multi-stage processing and separation of different steps, the overall load can be split up and a lot of different device inputs (vehicles, selling points, satisfaction check points) can be processed without noticeable delay. In our simulated tests, up to 250 devices sending information at intervals of 1 to 10 seconds were checked, without resulting in significant system load.
- Use of configurable processing blocks can improve a lot system flexibility and facilitate use of dashboards that present the output.
- Parallel processing and use of orchestration frameworks can help in maintaining sequence of different actions and trigger processing of inputs, only when it is necessary.

Conclusion

Consumer behavior analysis is extremely important in providing adequate, economically sustainable and efficient transportation services. Due to its significance for local community and important social aspects, municipal transportation system offers some special and unique products. In this report we have highlighted approaches and algorithms that can be used to collect information, study consumer behavior, process it in appropriate way and provide outputs that can support various stages of the municipal authorities decision-making processes.

Considering the goals and stages of the INNOAIR project, our suggestions can help greatly in planning green transportation and adapting it to meet in full consumer needs. There are several important characteristics of consumer behavior analysis presented in the report:

- First of all, it is considered as a continuous process that should provide updates with various frequency – starting from real-time monitoring of behavior to conducting periodic in-depth surveys of attitude toward public transportation system.
- Consumer behavior can be studied from different points of view, but to gain full insight on what drives people in choosing different transportation options, we need to combine data inputs from various sources and of various types.
- Machine learning algorithms can support the analysis, but only when properly combined with expert knowledge. This is particularly important for representing the outputs in a way that can be beneficial to people with different level of knowledge in technical details.
- Flexibility and adaptability are important characteristics in the scenarios regarding modes and services. In this way, the transport system will adjust to peoples' needs and to the way in which the urban environment is constantly changing in order to meet the needs and expectations of people.

We suggest a new, flexible and versatile software system, that relies exclusively on open source components and is capable of processing large amounts of information related to consumer satisfaction. This approach can provide a solid foundation for in-time feed of both technical and management information that improves operation of existing transportation networks and facilitates new approaches like those at the hearth of INNOAIR project.

Acknowledgement

This research was supported by UIA05-202 "INNOAIR - Innovative demand responsive green public transportation for cleaner air in urban environment", funded by the European Union initiative - Urban Innovative Actions. (UIA)."

References

[1] E. C. f. Standardization, "European Standard, EN 13816," 2002. [Online]. Available: https://tpbi.ro/file/2021/02/EN-13816-standard-Service-Quality-Definition-Targeting-and-Measurement-EU-2002.pdf.

[2] G. Mazzulla and L. Eboli, "A service quality experimental measure for public transport," *European Transport,* vol. 34, pp. 42-53, 2006.

[3] L. H. Yaya, M. F. Fortià, C. S. Canals and F. Marimon, "ervice quality assessment of public transport and the implication role of demographic characteristics," *Public Transport,* vol. 7, no. 3, pp. 409-428, 2015.

[4] A. Parasuraman, V. Zeithaml and L. Berry, "SERVQUAL: a multiple-item scale for measuring consumer perceptions of service quality," *Retailing: critical concepts,* vol. 64, no. 1, p. 140, 2002.

[5] J. J. Cronin Jr and S. A. Taylor, "SERVPERF versus SERVQUAL: reconciling performance-based and perceptions-minus-expectations measurement of service quality," *Journal of marketing,* vol. 58, no. 1, pp. 125-131, 1994.

[6] C. Fornell, M. D. Johnson, E. W. Anderson, J. Cha and B. E. Bryant, "The American customer satisfaction index: nature, purpose, and findings," *Journal of marketing,* vol. 60, no. 4, pp. 7-18, 1996.

[7] J. De Oña, R. De Oña, L. Eboli and G. Mazzulla, "Perceived service quality in bus transit service: a structural equation approach," *Transport policy,* vol. 29, pp. 219-226, 2013.

[8] H. M. Clark, "Who Rides Public Transportation," APTA- American Public Transportation, https://www.apta.com/wp-content/uploads/Resources/resources/reportsandpublications/Documents/APTA-Who-Rides-Public-Transportation-2017.pdf, 2017.

[9] Gogola, M., Sitanyiová, D., Černický, L., Veterník. M., "NEW DEMAND PATTERNS FOR PUBLIC TRANSPORT DUE TO DEMOGRAHIC CHANGE," INTERREG CENTRAL EUROPE, https://www.interreg-central.eu/Content.Node/working-paper--New-demand-patterns-for-public-transport-due-.pdf, 2018.

[10] Şimşekoğlu, Ö., Nordfjærn,T., Rundmo, T., "The role of attitudes, transport priorities, and car use habit for travel mode use and intentions to use public transportation in an urban Norwegian public," *Transport Policy, Elsiever,* pp. 113-120, 2015.

[11] NSI, "Population by districts, age, place of residence and sex," 2022.

[12] NSI, "Population projections by districts and sex," 2018.

[13] NSI, "Projected age dependency ratio by districts and sex until 2080," 2018.

[14] Statista, "Average number of persons per household in selected European countries in 2020," 2022.

[15] Eurostat, "2021," Stocj of vehicles by category and NUTS 2 regions.

[16] Sofiaplan, "Infographics," 2021.

[17] J. Bastiaanssen, D. Johnson and K. Lucas, "Does better job accessibility help people gain employment? The role of public transport in Great Britain," *Urban Studies,* vol. 2021, pp. 1-22, 2021.

[18] Y. Han, W. Li, S. Wei and T. Zhang, "Research on Passanger`s Travel Mode Choice Behavior Waiting at Bus Station Based on SEM-Logit Model," *Sustainability,* 1996.

[19] Y. Godwin, C. Cottrill and N. J., "Understanding factors influencing public transport passengers pre-travel information-seeking behaviour," *Public Transport,* pp. 135-158, 2019.

[20] B. Sweeney, "An Analysis of the Role of Real Time Passenger Information on Bus Users in a European City: The Case of Dublin, Ireland," ITRN , Dublin, 2012.

[21] J. Legal, T. Meyer and A. Csillik, "Goal Primign, Public Transportation Habit and Travel Mode Selection: The Moderating Role of Trait Mindfulness," TRANSPORT RES F-TRAF, 2016.

[22] S. Sumaedi, I. Bakti and M. Yarmen, "The Empirical Study of Public Transport Passengers Behavioral Intentions: The Roles of Service Quality, Perceived Sacrifice, Perceived Value, and Satisfaction (Case Study: Paratransit Passengers in Jakarta, Indonesia)," *International Journal for Traffic and Transport Engineering,* pp. 83-97, 2012.

[23] G. Beirao and J. A. S. Cabral, "Understanding attitudes towards public transport and private car: A qualitative study," *Transport Policy,* vol. 14, no. 6, pp. 478-489, 2007.

[24] L. Steg, "Car use: lust and must. Instrumental, symbolic and affective motives for car use," *Transportation Research: Part A: Policy and Practice 39,* vol. 2, no. 3, pp. 147-162, 2005.

[25] M. R. Solomon, Consumer Behavior: Buying, Having, and Being, Pearson Prentice Hall, 2004.

[26] Y. Gaoa, S. Rasoulib, H. Timmermansb and Y. Wang, "Trip stage satisfaction of public transport users: A reference-based model incorporating trip attributes, perceived service quality, psychological disposition and difference tolerance," *Transportation Research Part A,* vol. 118, p. 773, 2018.

[27] J. De Vos, P. L. Mokhtarian, T. Schwanen, V. Van Acker and F. Witlox, "“Travel mode choice and travel satisfaction: bridging the gap between decision utility and experienced utility," *Transportation,* vol. 43, no. 5, p. 771–796, 2016.

[28] A. Carrel, R. G. Mishalani, R. Sengupta and J. L. Walker, "In pursuit of the happy transit rider: dissecting satisfaction using daily surveys and tracking data," *J. Intell. Transp. Syst.,* vol. 20, no. 4, p. 345–362, 2016.

[29] Apache Software Foundation, "Apache NiFi," 11 11 2021. [Online]. Available: https://nifi.apache.org/. [Accessed 11 11 2021].

[30] influxdata, "InfluxDB Telegraf," 15 11 2021. [Online]. Available: https://www.influxdata.com/time-series-platform/telegraf/. [Accessed 15 11 2021].

[31] Apache Software Foundation, "Apache Airflow," 15 11 2021. [Online]. Available: https://airflow.apache.org/. [Accessed 15 11 2021].

[32] Apache Software Foundation, "Apache Spark MLib," 18 11 2021. [Online]. Available: https://spark.apache.org/mllib/. [Accessed 18 11 2021].

[33] Grafana Labs, "Grafana," 22 11 2021. [Online]. Available: https://grafana.com/. [Accessed 22 11 2021].