PASSENGER DEMAND PREDICTION FOR METRO STATIONS USING PROBABILISTIC MODEL

AIM To explore how (Probabilistic Model) Gaussian, Binomial & Log linear models can predict Passenger demand at metro stations.

RESEARCH QUESTION Out of these three models which works best for metropolitan cities

TASK OBJECTIVE 1
Identification and selection of parameters influencing the passenger demand

OBJECTIVE 2
Data collection of the parameters & Site selection

OBJECTIVE 3
Model building & comparison and assessing of parameters

OBJECTIVE 4
To validate the model in other cities and Suggest a way forward

NEED OF STUDY

Demand Modeling methods used in DPR's in India, dates back 1960. Certain limitations imposed by models are:

Four Step Travel Demand Modeling (1960)	**Oversimplifies** the complex factors influencing transportation demand
Gravity Models (1960)	**Overlooks** other significant factors, like land use patterns and travel cost
Disaggregate Choice Models (1970)	**Neglects** individual variations and heterogeneity in decision-making processes.

Due to this the inaccurate estimation of Passenger Demand

Projected, actual ridership and shortfall in 2019-2020 Of Delhi MRTS

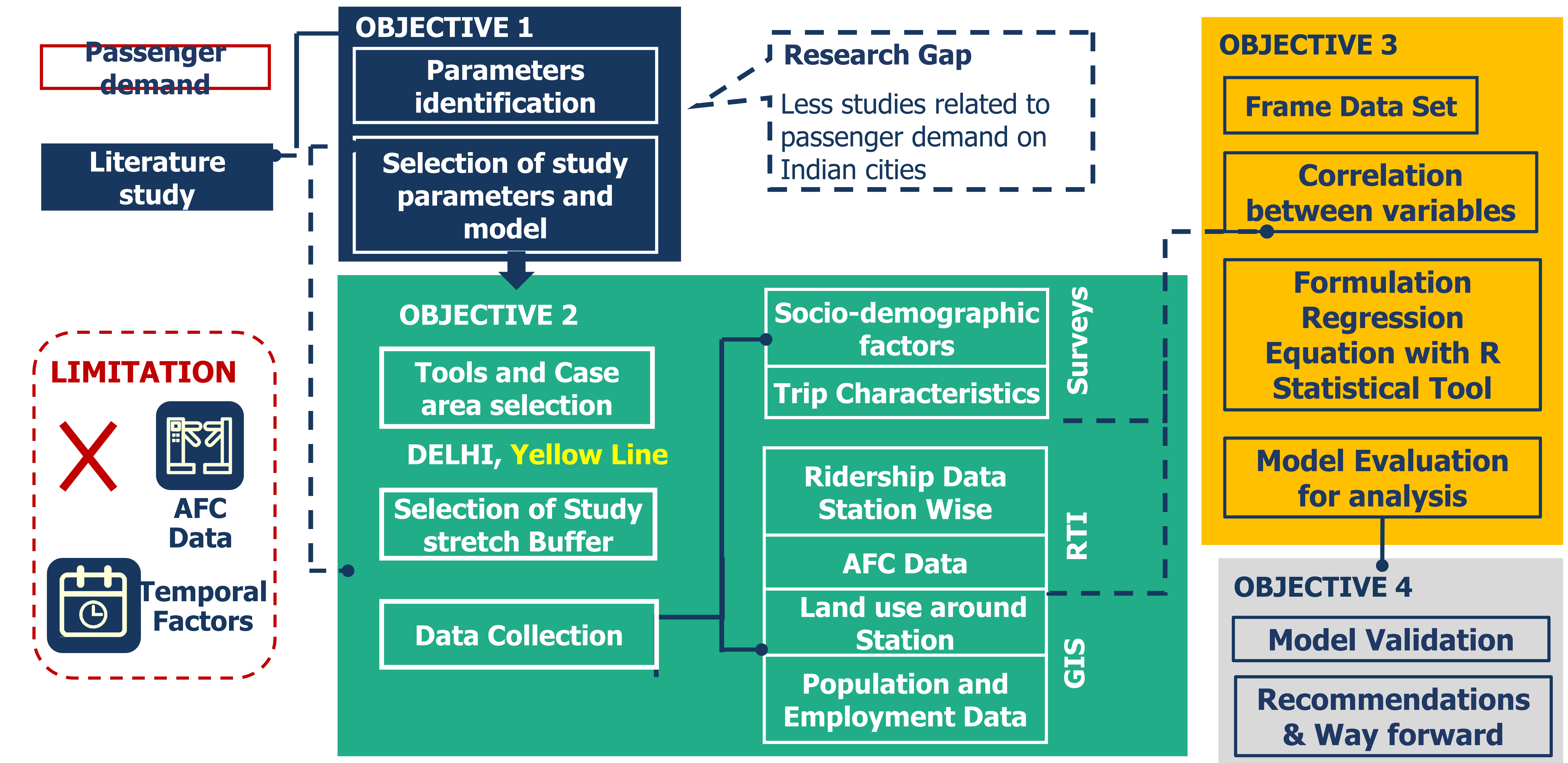
Phase/Line	Estimated Ridership	Actual Ridership	Percentage Shortfall
Phase-I (DPR, 1995)	31.85 lakh	6.62 lakh	79%
Phase-I (Revised, 2003)	22.60 lakh	6.62 lakh	71%
Phase-II (Airport Line)	42,500	17,794	58%
Phase-I, II, and III (2019-20)	53.47 lakh	27.79 lakh	48%

Source: CAG Audit Report no. 11 of 2021

Projected, actual ridership and shortfall in 2021 of other MRTS

Phase/Line	Estimated Ridership	Actual Ridership	Percentage Shortfall
Phase-I (DPR, 2005) Mumbai metro Line 1	6.7 lakh	4.5 lakh	33%
Phase-I (DPR, 2003) Bengaluru metro	16.1 lakh	4.5 lakh	72%

RESEARCH FRAMEWORK



SCOPE

- The following study will solely focus on **daily passenger demand prediction** for metro, as AFC data was not provided by the DMRC for an hourly passenger demand analysis.
- It is solely focused on the surface area of **land use around the metro station**, without considering the intensity of land use

EXPECTED OUTCOME

- To provide a quantitative **understanding of the relationship between selected variables and passenger demand**.
- To suggest changes in demand modeling procedures followed in existing DPRs.

PASSENGER DEMAND PREDICTION FOR METRO STATIONS USING PROBABILISTIC MODEL

Literature study

METRO STATION PASSENGER DEMAND

Region	Method	Topic	Literature
AMERICA	Multiple Regression	Passenger Demand	Cerevro et al. 1997; Kuby et al. 2004
EUROPE	Multiple Regression, GWR	Transit Ridership	Gutierezz et al. 2011; Cardazo et al. 2014
ASIA	Multiple Regression, ANN	Passenger demand	Sung H et al. 2011; Lin et al. 2020; Yang et al. 2021

Built Environment

- Residential Land
- Commercial Land
- Mixed Use Land
- Population Density
- Employment Density
- Road Density

Station Characteristics

Factor	+	−
Parking Facility	Convenience	Limited Capacity
Transfer Station	Connectivity	Congestion
Station Spacing	Accessibility	Inefficiency

Socio Economic

- Employment
- Gender
- Age Group
- Population
- Households
- Income group

Trip Characteristics

Factor	+	−
Trip Fare	Affordability	Costliness
Journey Time	Longer	Shorter
Trip Frequency	More Frequent	Less Frequent

Temporal Factors

Factor	+	−
Type of Day	Weekday	Weekend
Time	Peak Hour	Non Peak
Weather	Temperature	Precipitation

DETERMINISTIC MODEL APPROACH

It does not take into account any randomness or uncertainty in the data

Optimization Models	Linear Programming
Simulation Model	Traffic Simulation
Rule Based Model	Expert Systems
Descriptive Model	Factor Analysis

PROBABILISTIC MODEL APPROACH

Mathematical framework for representing uncertain quantities and their relationships

Parametric	**Regression, Log linear, Negative Binomial**
Non-Parametric	Kernel Density estimation, Decision trees
Time Series	ARMA, ARIMA
Machine Learning	ANN, SVR, Random Forests

Parametric

- Data follows a specific distribution
- Outcome can be continuous or count variable
- Variables having multiplicative relationship

Advantages over other models

- **Flexibility**: Accommodate a wide range of functional forms and distributions, for better representation of the data patterns.
- **Generalizability**: It can generalize well to new data, allowing for reliable predictions in similar contexts.
- **Interpretability**: Provides interpretable coefficients for evaluating the impact of predictor variables on the outcome.

Literature	Shortcomings of 4S-TDFd in DPRs	Description
N. Oppenheim 1995	Sequential nature of the procedure	The step-by-step **approach lacks a unifying rationale**, making it difficult to understand and communicate to decision-makers.
Y. Gu et al. 2004	Aggregation of behavior	Aggregate models **cannot predict individual traveler behavior,** relying on macro-level assumptions.
Donnelly R. et al. 2004	Deterministic nature of the models	Models are **mathematical rather than simulation-based,** limiting their ability to simulate real-world scenarios.
Boyce D. et al. 2002	Iterative nature of the process	**Travel costs are not in equilibrium condition,** requiring iterative feedback to approach network equilibrium.
C. A. Flaherty,1997	Approach to prediction	**The focus on trend extrapolation rather than a rational goal** limits the ability to modify present trends.
R Johnston, 2004	Integrated land-use and transportation models	**Neglecting the feedback between transportation and land use** hinders the support for land use policies.
V. R. Vuchic 2005,	The effects of congestion	**Congestion effects and demand externalities are not adequately considered,** affecting the precision of travel demand estimates.
Donnelly R. et al. 2004	Input data issues	**Heavy reliance on limited household travel survey and census** data affects the development and calibration of complex models.

OBJECTIVE 1: IMPACT OF PARAMETERS AND MODEL SELECTION

PASSENGER DEMAND PREDICTION FOR METRO STATIONS USING PROBABILISTIC MODEL

Q-GIS

R Studio

Tools selection

STEP-1 → **Problem Definition**

To predict correct estimates of Passenger Demand of an MRTS

STEP-2 → **Parameters Identification** → **Data Collection and Analyzing**

STEP-3 → **Correlation between variables**

STEP-4 → **Model Specification for analysis**

Gaussian Linear Model: It assumes that the response variable (passenger demand) is normally distributed

Negative Binomial: It is appropriate when the variance in passenger demand is not constant across different levels of the predictor i.e. over dispersion and response variable is a count variable

Log Linear Regression: It assumes that the response variable (passenger demand) follows a Poisson distribution. It is used when response variable is a count variable, and the variance is equal to the mean

STEP-5 → **Model Evaluation for analysis**

Input Variables in all three Regression Equation

Using measures R^2, Akaike Information Criteria, MAPE values

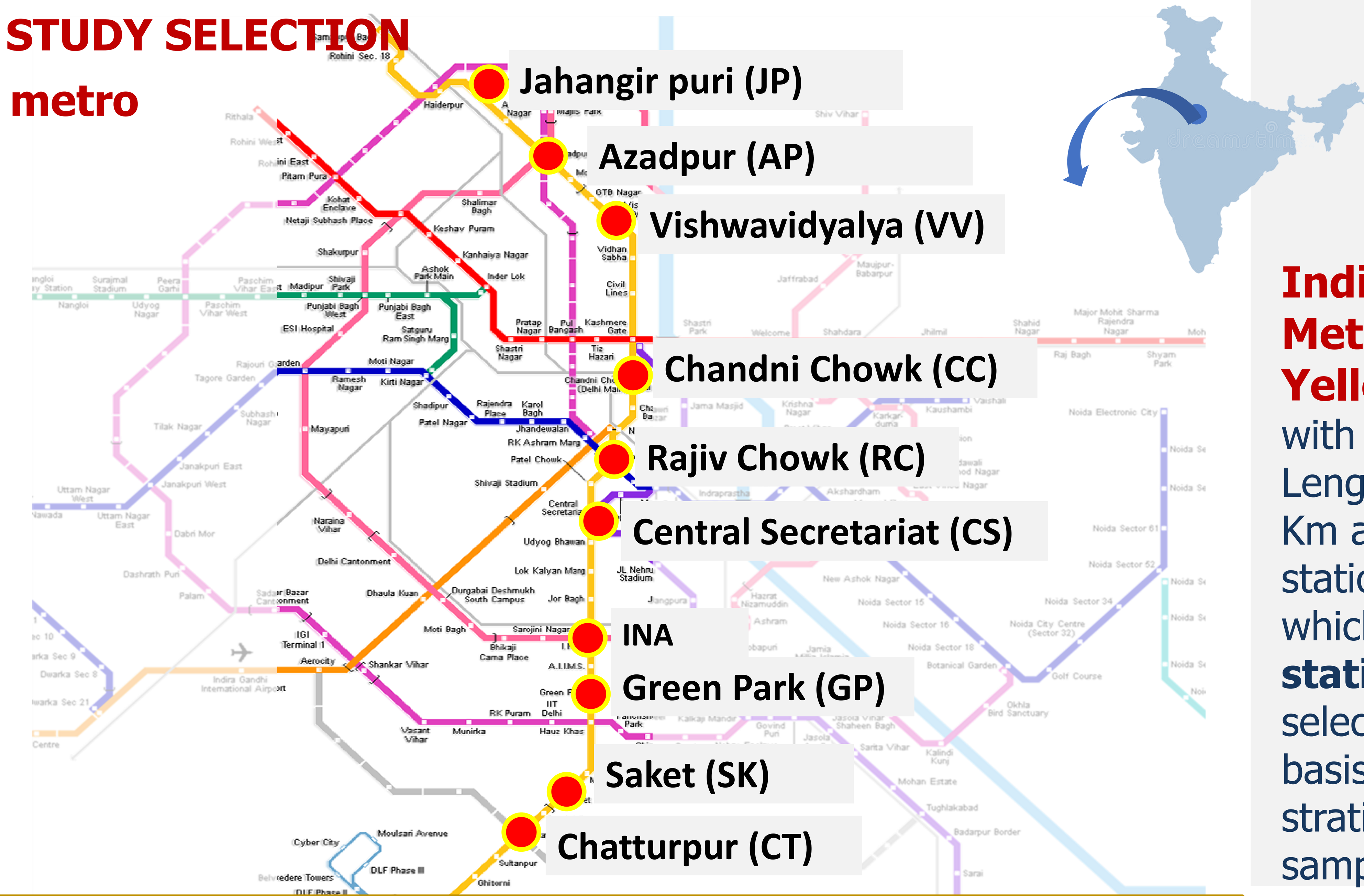
STEP-6 → **Model Validation**

Testing model on new data

Comparison of all three models and efficiency evaluation

CASE STUDY SELECTION

Delhi metro Map



India, Delhi, Metro Yellow line with Total Length 49.31 Km and 37 stations out of which **10 stations** were selected on the basis of stratified sampling.

Line Selection: Ridership

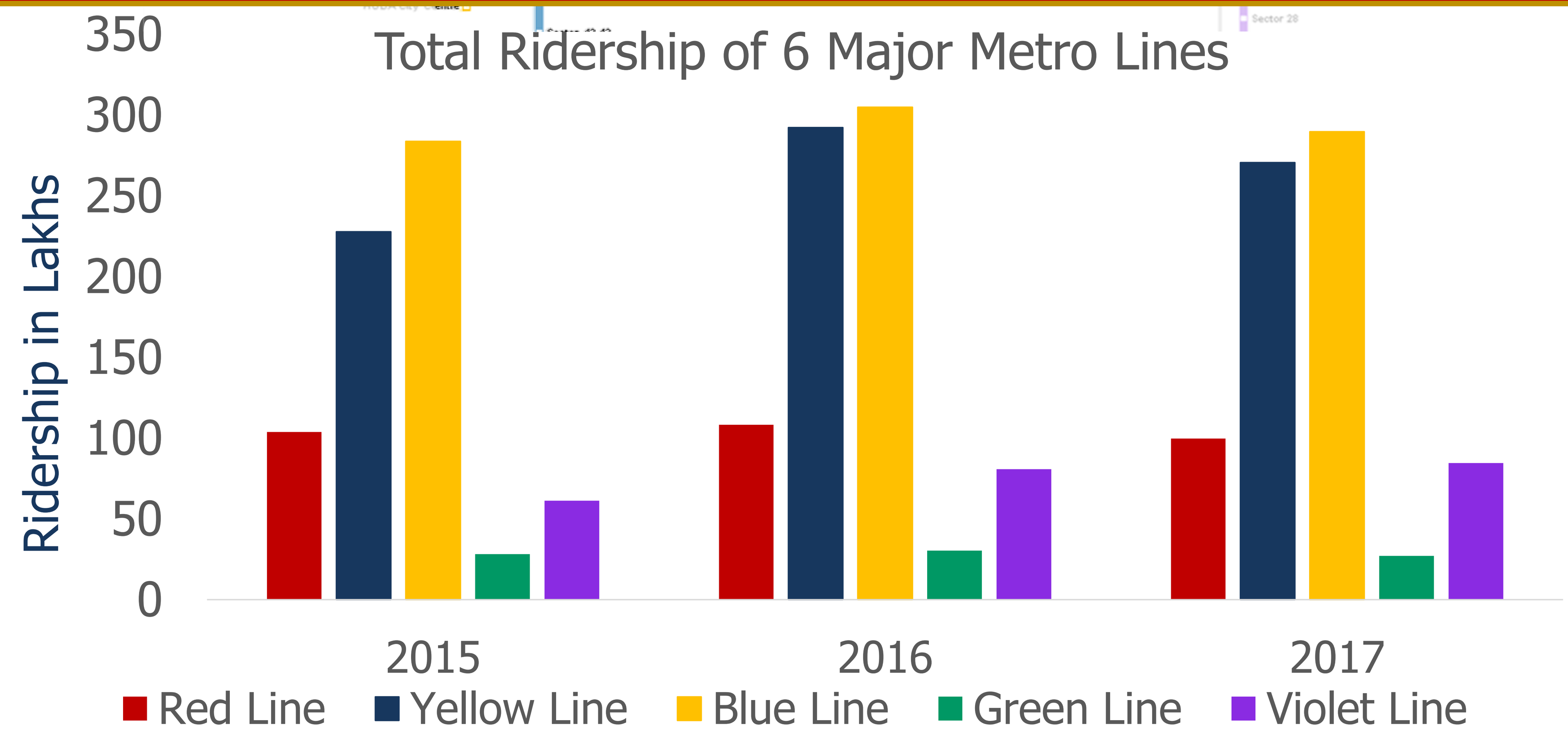


Figure: Ridership details of major metro lines in Delhi . Source: DMRC

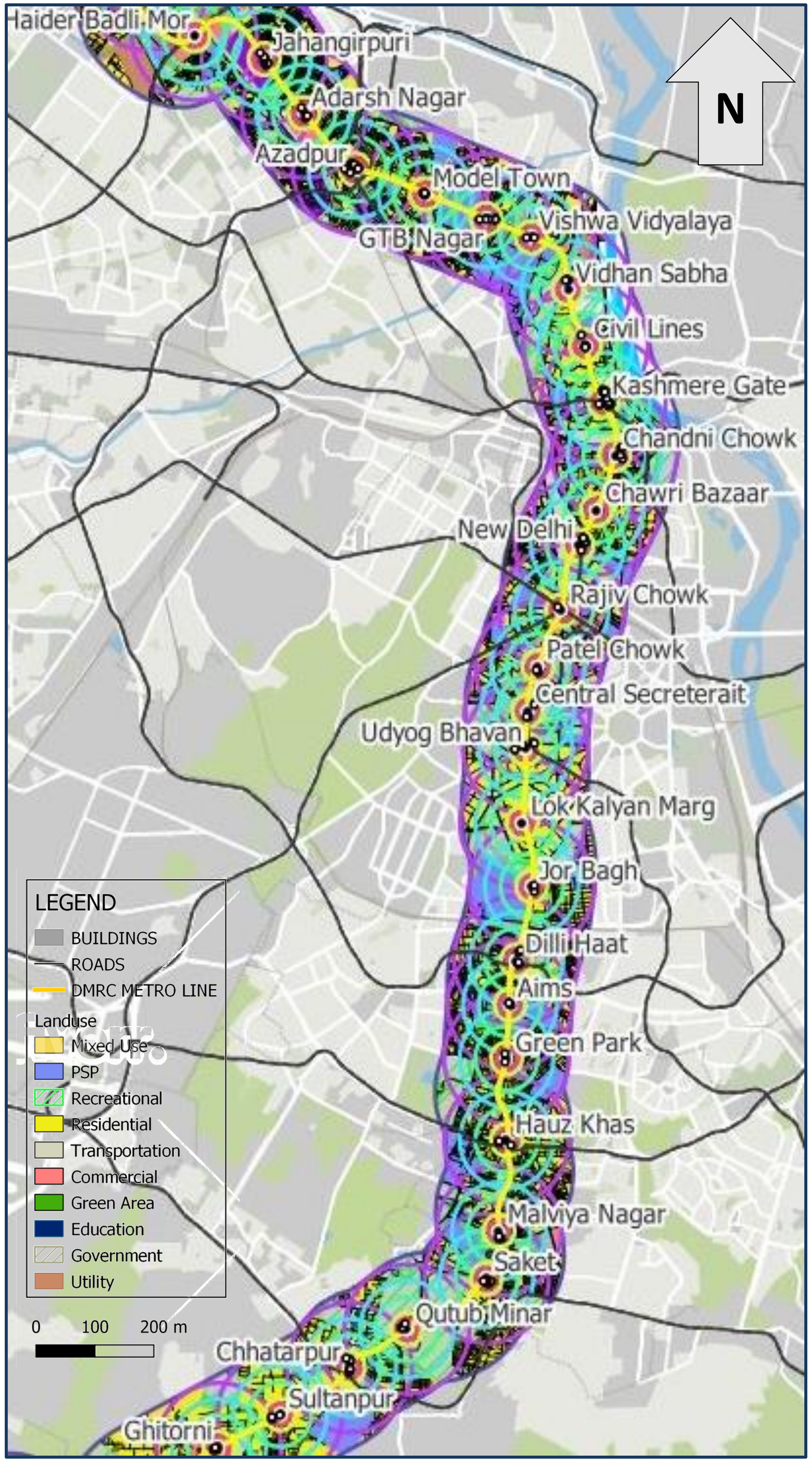
Due to **High Ridership, Best Service frequency and Congestion faced during Peak Hours, Yellow Line** has been selected.

Line Selection: Capacity and LOS

Line	Type of train	PHPDT (July 16)	Headway (sec)	Trains/ Hour	Total	Total Cars	Average no. of cars/train	No. of cars in Peak Hours	Pax/ Sq.m.	LOS
Red	RS-1	22528	195	18.5	29	136	4.7	87	5.1	D
Yellow	RS-1&2	55500	133	27.1	60	426	7.1	183	6.1	E
Blue	RS-1&2	45935	150	24	71	476	6.7	161	5.7	D
Green	RS-3	10839	228	15.8	23	94	4.1	65	3.4	C
Violet	RS-3	20229	200	18	44	264	6	108	3.9	C

Table: Standing capacity and LOS analysis of major metro lines in Delhi . Source: Sarkar and Jain

OBJECTIVE 2: TOOLS, CASE STATION AND LINE SELECTION



PROCESS-1: Identification of ideal service range of metro station

CIRCULAR BUFFER: 500 M

CIRCULAR BUFFER: 750 M

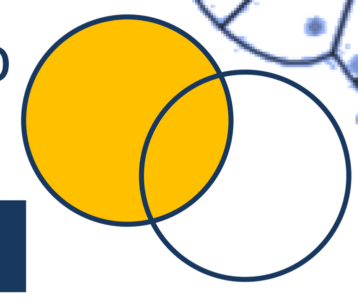
CIRCULAR BUFFER: 1000 M

"The transit industry widely applies the 400-meter (0.25-mile) and 800-meter (0.5-mile) rules of thumb when estimating service areas around bus and rail stations." — El-Geneidy et al. (2014)

PROCESS-2: Processing method of overlapping area of case metro stations.

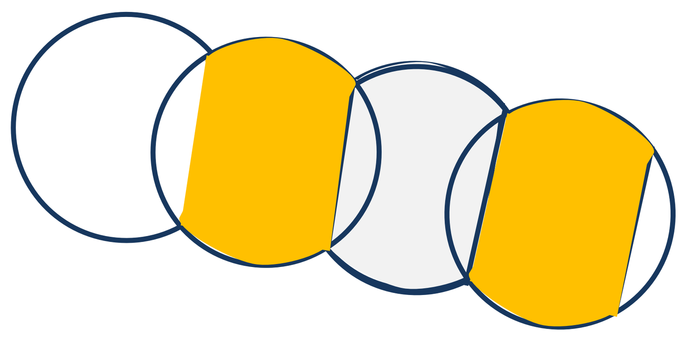
Naïve method

- Overlapping area will be counted into all buffers.



Equal division

- Divides the overlapping area by the number of overlapping buffers



Thiessen polygon

- Voronoi Dig. It creates a network of polygons that collectively cover the entire area of interest.

Note: We have used Naïve Method

QGIS Tool used to marked Land use within Buffer. Each variable is (Independent)

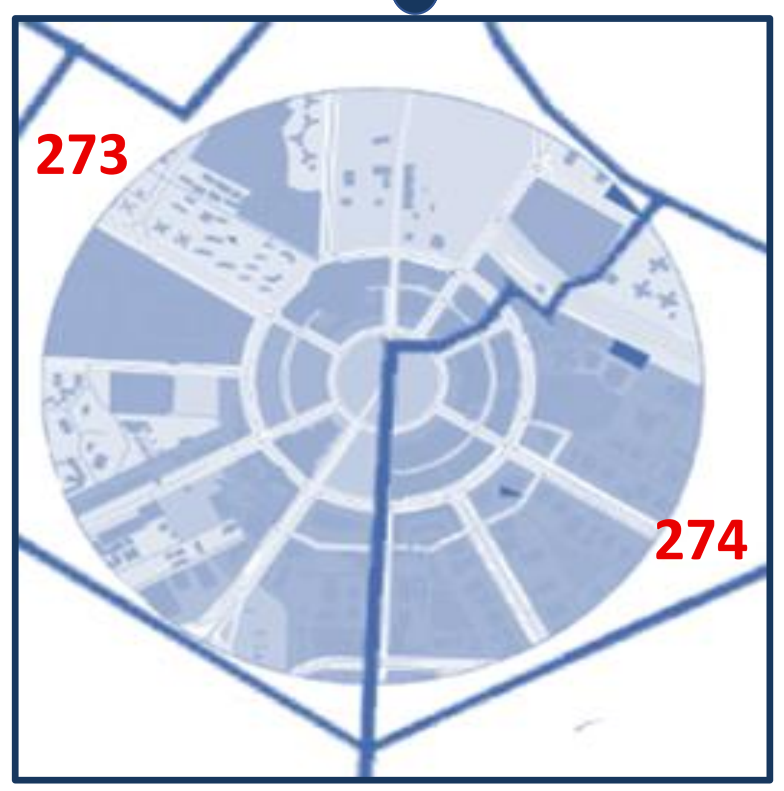


Figure: Shows the Land use of Buffer Area of 750 m Around Rajiv Chowk(RC) Station. Buffer Zone is overlapping TAZ 273 & 274

Using the Weighted Average method, we can calculate the population and employment density for the buffer area as follows:

1	Population density of TAZ	PD = POP / TA
2	Area of overlap between TAZ & Buffer	BA
3	Weighted factor of TAZ No. = Area of overlap / Total area of TAZ	WF = BA / TA
4	Population density of buffer zone for TAZ	PD_BZ = PD * WF

Example of calculation for TAZ 273 of Rajiv Chowk metro station

1	Population density of TAZ 273	295.92 persons/hectare
2	Area of overlap between TAZ 273 and buffer zone	108.1 hectares
3	Weight of TAZ 273 = Area of overlap / Total area of TAZ 273	108.1 / 184.1 = 0.5874
4	Population density of buffer zone for TAZ 273 =PD * WF	295.91 * 0.5874 = 173.7 persons/hectare

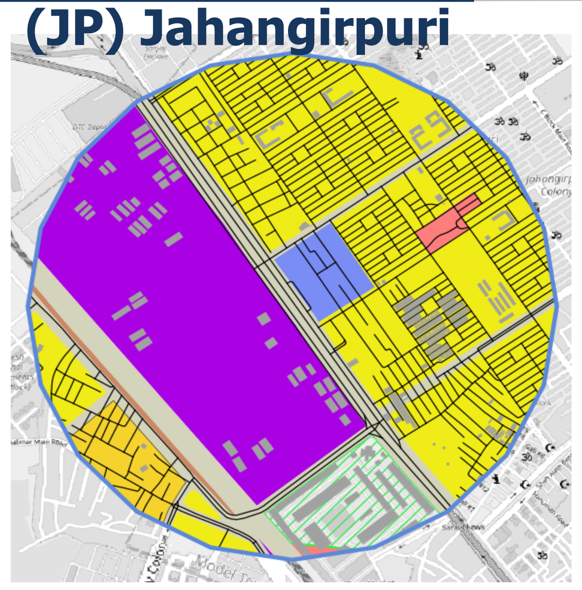
OBJECTIVE 2: ASSESING THE PARAMETERS WITH IN SERVICE RANGE

Selected Parameters

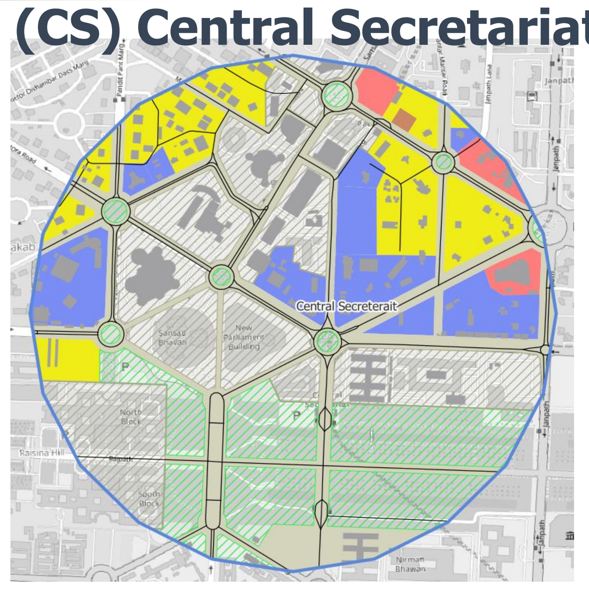
To focus the analysis, and gain deeper insights into the specific aspects, three parameter classes were taken into consideration, namely, Socio-economic variables, Station characteristics and Built environment characteristics. The subset of these parameters, their description, variable type and their mode of collection is as shown

Variables	Description	Variable Type	Mode of Collection
Socioeconomic Variables			
POP	Population Count	Continuous	RITES, 2010
EMP	No of Workers	Continuous	RITES, 2010
AGE	Age group of people utilizing the station	Discrete Ordinal	Primary survey through Interview/ Questionnaire
INC	Income level of people using the station	Discrete Ordinal	
Built Environment Variables			
PD	Population Density (persons/hectare)	Continuous	GIS Tool
EMD	Employment Density (persons/hectare)	Continuous	
RES	Area of Residential Land (hectares)	Continuous	GIS Tool
COM	Area of Commercial Land (hectares)	Continuous	
MIX	Area of Mixed use Land (hectares)	Continuous	
REC	Area of Recreational Land (hectares)	Continuous	
EDU	Area of Educational Land (hectares)	Continuous	
PSP	Area of Public-semipublic (hectares)	Continuous	
EDU	Area of Educational Land (hectares)	Continuous	
Station Characteristics			
TR	Station is Transfer =1; Otherwise = 0	Binary	Through DMRC

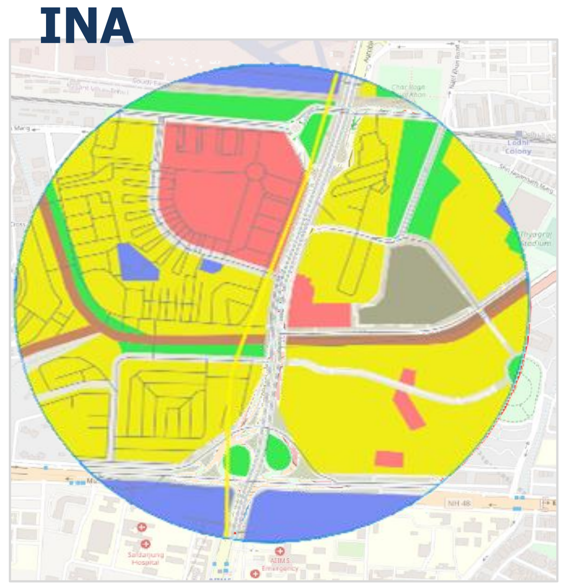
(JP) Jahangirpuri



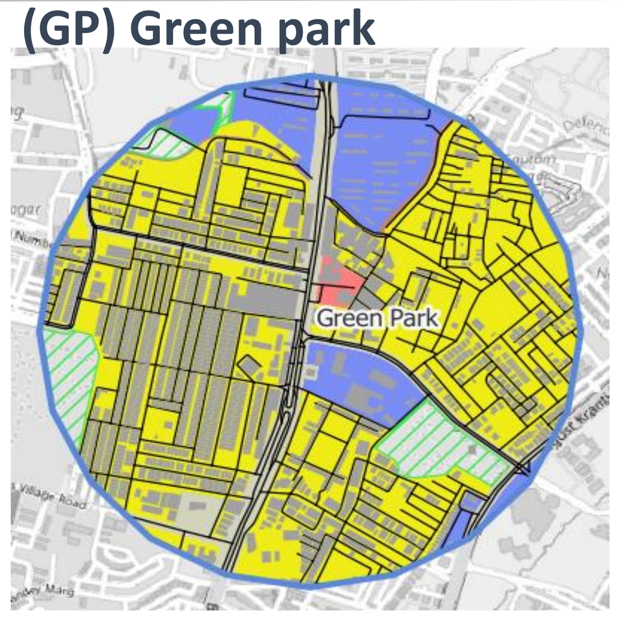
(CS) Central Secretariat



INA

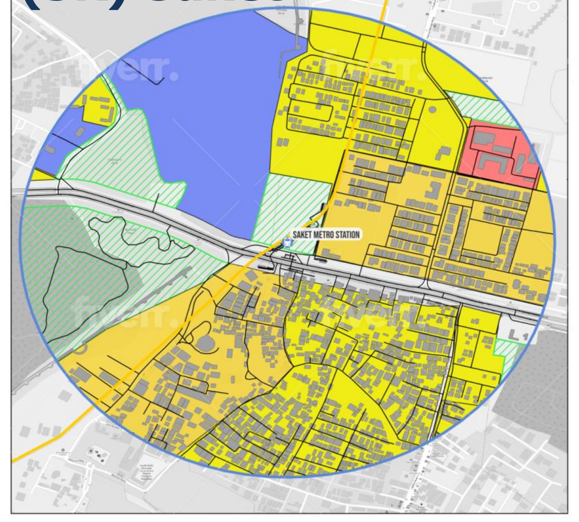


(GP) Green park

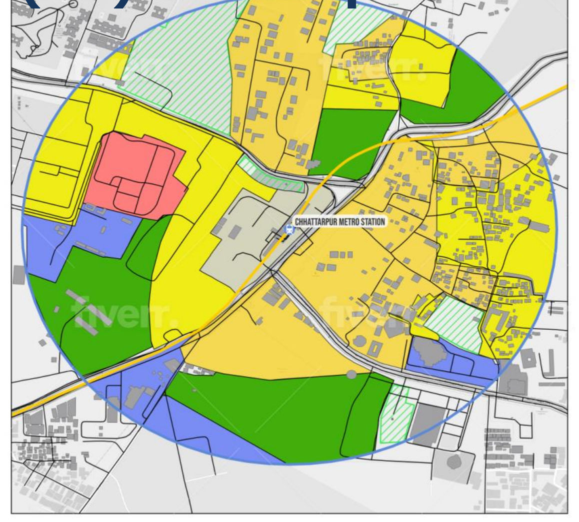


(AZ) Azadpur

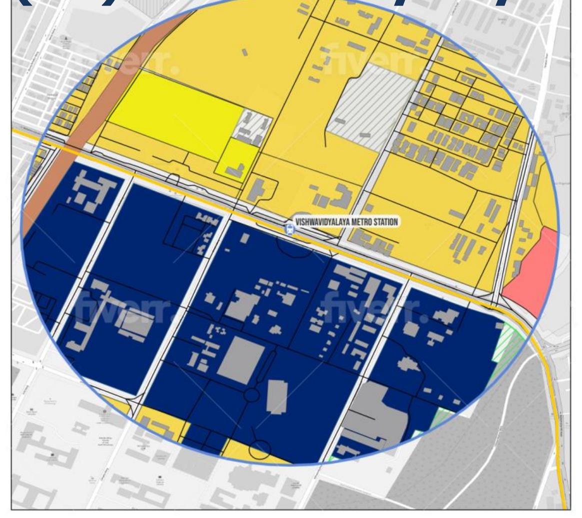
(SK) Saket



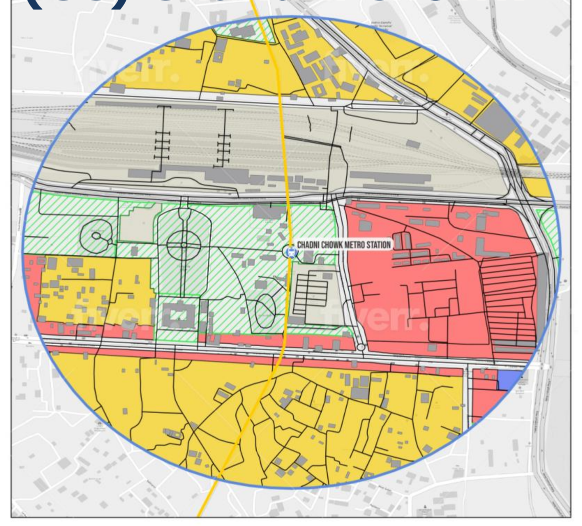
(CH) Chhatarpur



(VV) Vishwavidyalaya



(CC) Chandni chowk



(RC) RAJIV CHOWK

Land Use of Selected metro stations of Yellow line, Delhi

Land use within 750 m of buffer area of selected Metro Stations.

LEGEND
- BUILDINGS
- ROADS
- DMRC METRO LINE

Landuse
- Mixed Use
- PSP
- Recreational
- Residential
- Transportation
- Commercial
- Green Area
- Education
- Government
- Utility

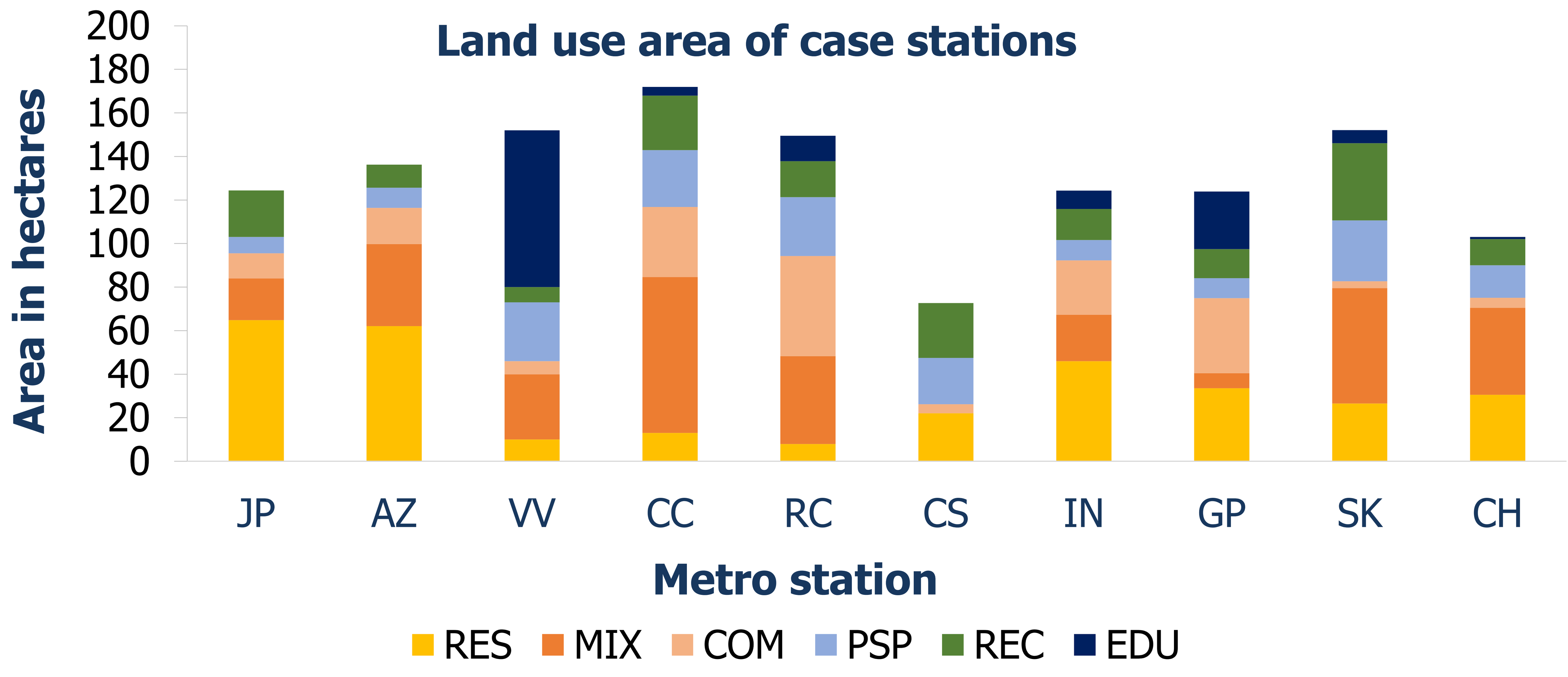
Data set comprising of Land use Area in Hectares, Population and Employment, Age and Income of Passengers

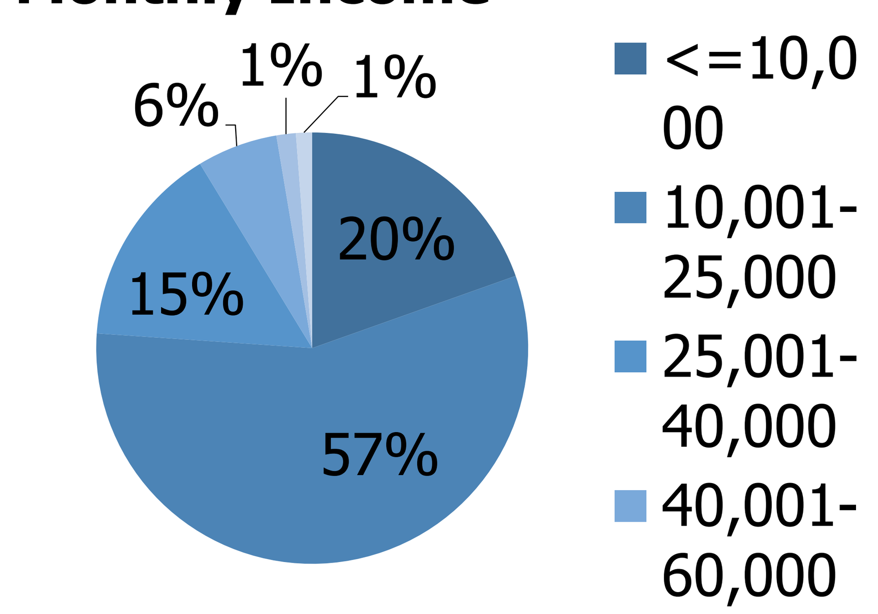
	PF	RES	MIX	COM	PSP	REC	EDU	PD	EMD	TR	AGE	POP	EMP	INC
JP	17345	64.8	19.2	11.5	7.5	21.4	0	168.45	53.9	0	3	28149	9014	3
AZ	7463	62.1	37.7	16.6	9.2	10.6	0	113.05	57.5	1	2	19185	9762	2
VV	19655	10	30	6	27	7	72	36.7	14.4	0	2	6268	2457	1
CC	46944	13	71.6	32.2	26.1	25	4	415.03	403.5	1	3	74913	72824	2
RC	40780	7.9	40.4	46	27	16.5	11.7	27.04	454.7	1	4	4750	79898	3
CS	11514	22	0	4.2	21.3	25.2	0	32.9	6.5	1	3	5593	1104	3
INA	9628	46	21.3	25	9.3	14.3	8.4	76.66	32.7	1	3	12648	5395	2
GP	16396	33.5	7	34.4	9.2	13.3	26.5	45.5	20.8	0	4	7800	3562	2
SK	29370	26.5	53	3.2	27.9	35.5	6	19.24	8.6	0	2	3378	1514	2
CP	24369	30.5	40	4.6	14.9	12	1	7.35	5.6	0	2	1270	974	2

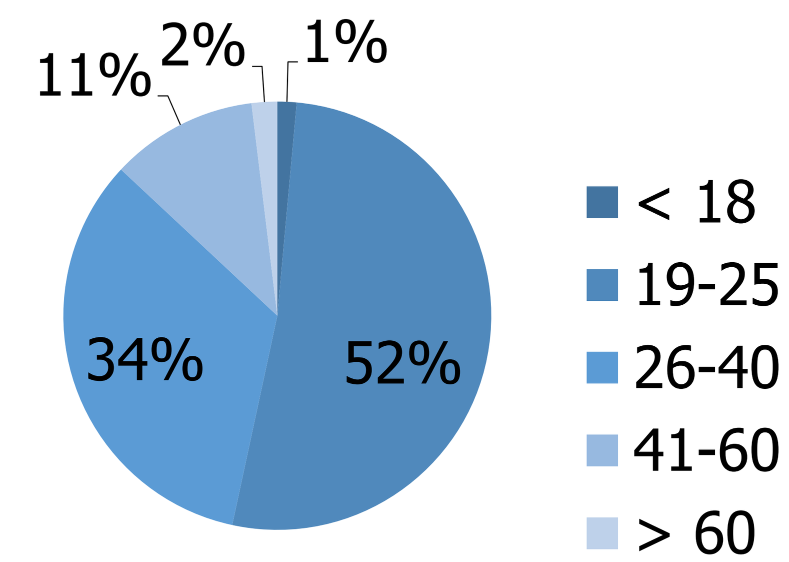
OBJECTIVE 2: PARAMETERS SELECTION AND ASSESSING THEM

MODEL BUILDING AND COMPARISON

STEP-1: Formulation of data set comprising of Land use Area in Hectares, Population and Employment, Age and Income of Passengers



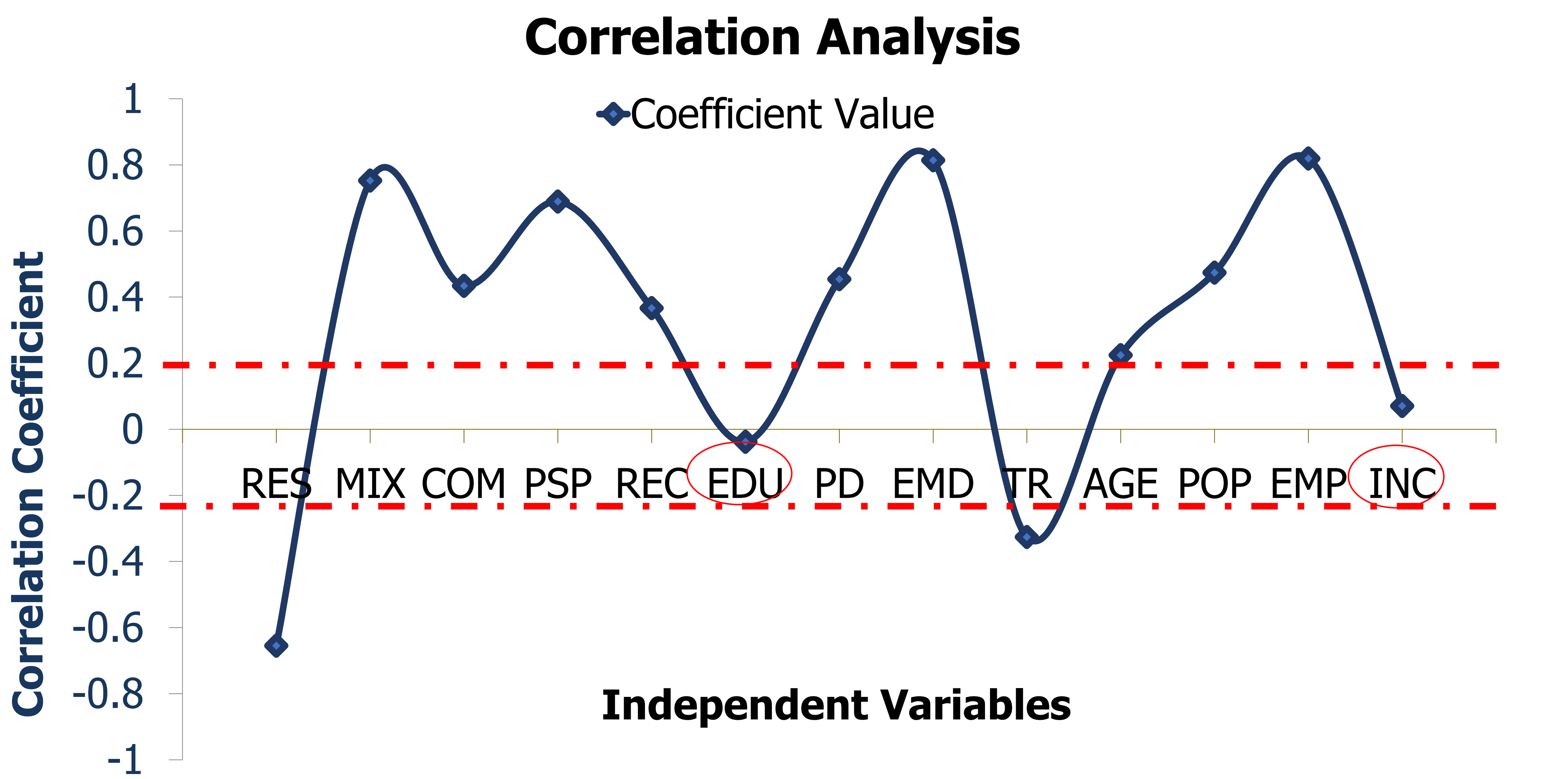




Passenger Age

PF	RES	MIX	COM	PSP	REC	PD	EMD	TR	AGE	POP	EMP	EDU	INC
17345	64.8	19.2	11.5	7.5	21.4	168.45	53.9	0	3	28149	9014	0	3

STEP-2: Checking the Correlation between the variables and Eliminating the variables which are not statistically significant



The correlation between variables was checked and a bivariate correlation matrix was generated. **The Pearson Coefficient** were used to test the correlation between the variables and variables having value **< 0.25** are eliminated as they are weakly correlated with PF

SYMBOL	NOTATION
PF	Passenger demand
RES	Residential Land
COM	Commercial Land
EDU	Educational Land
REC	Recreational Land
PSP	Public-Semi Public
POP	Population Count
EMP	Employment Count
JP	Jahangirpuri
AZ	Azadpur Station
VV	Vishwavidyalaya
CC	Chandni Chowk
RC	Rajiv Chowk
IN	Dilli Haat-INA
GP	Green Park
SK	Saket
CP	Chhatarpur

STEP-3: Using the R statistical software the relationship between the passenger demand and independent variables is obtained

	Command	Description
1	read.csv("Passenger demand.csv")	# Reading the data
2	corr_matrix <- cor(data)	# Carry out the Correlation between variables
3	write.csv(cor(data) , "cor_results.csv")	# Export the results to a CSV file
4	model_1 <- glm(PF ~ RES + COM + MIX + PSP + AGE + EMP, family = gaussian, data = data)	# Gaussian Linear Model
5	model_2 <- glm.nb(PF ~ RES + MIX + COM + PSP + AGE + PD + EMD + TR, data = data)	# Negative Binomial Regression
6	model_3 <- glm(PF ~ RES + MIX + COM + PSP + AGE + PD + EMD + TR, family=poisson(link = "log"), data = data)	# Log Linear Model
7	Summary(model)	# Find out the summary
8	write.csv(summary(model)$coefficients, "glm_results.csv")	# Export the results to a CSV file

STEP-4: Formulation of Regression Equation btw the variables

Equation

$$PF = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \ldots + \varepsilon$$

Where β_0 is a constant and ε is an error term

Where β_1 β_2 β_3 β_4 are the coefficients parameters

Where X_1 X_2 X_3 X_4 ..are the independent variables

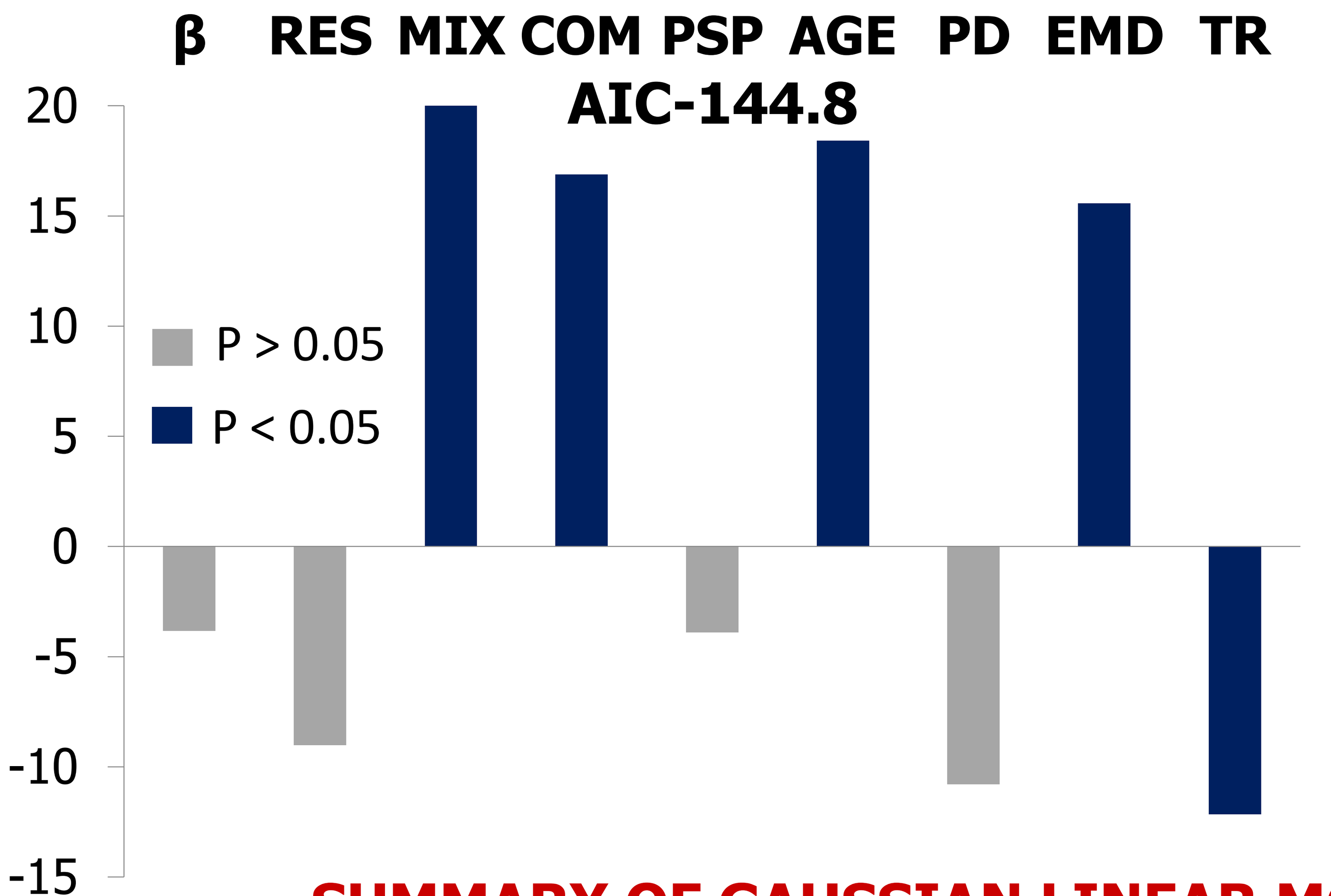
Interpreting the results in three models and Predicting Passenger demand

OBJECTIVE 3: FORMULATION OF PROBABILISTIC MODEL

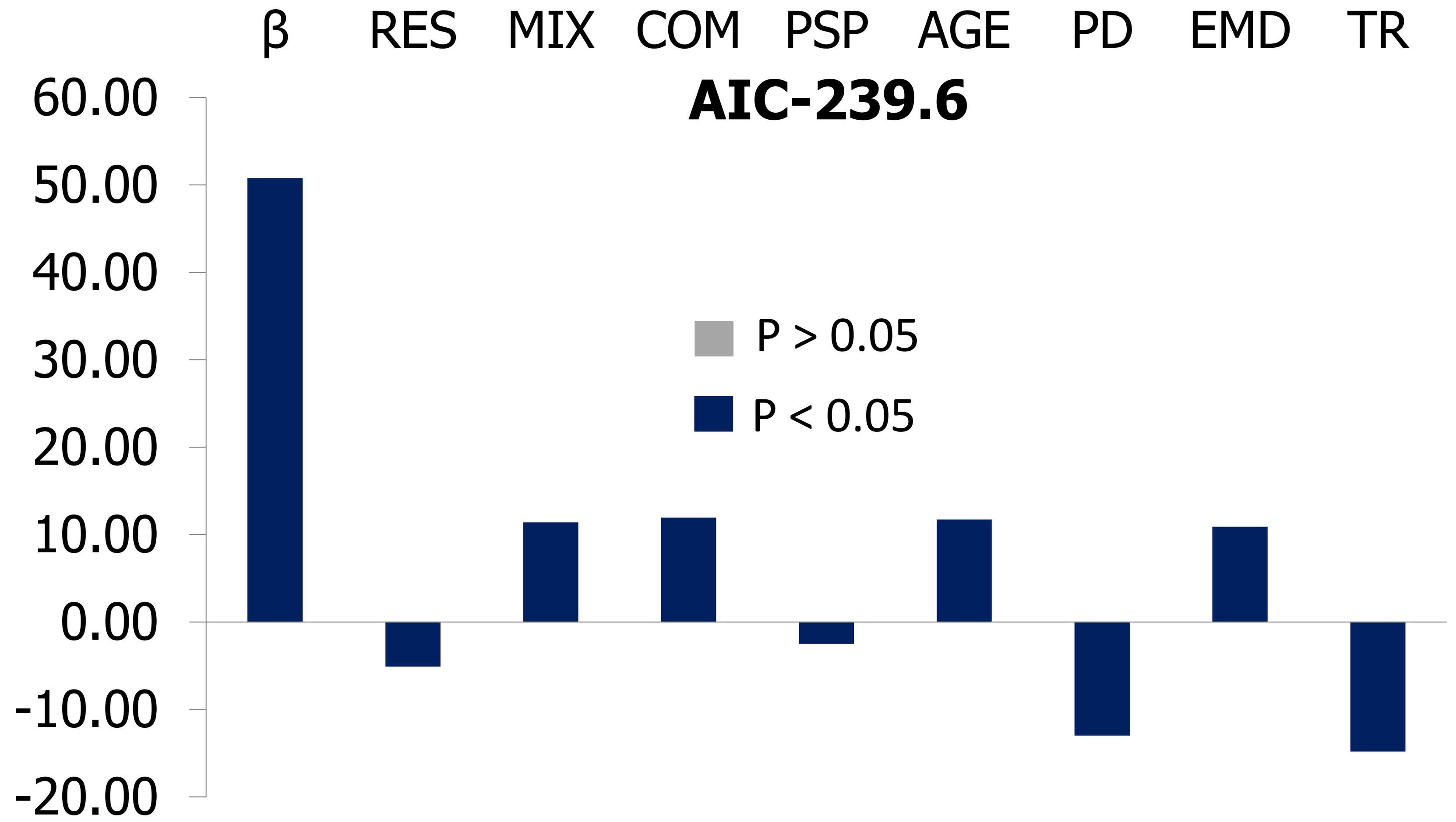
PASSENGER DEMAND PREDICTION FOR METRO STATIONS USING PROBABILISTIC MODEL

MODEL BUILDING AND COMPARISON

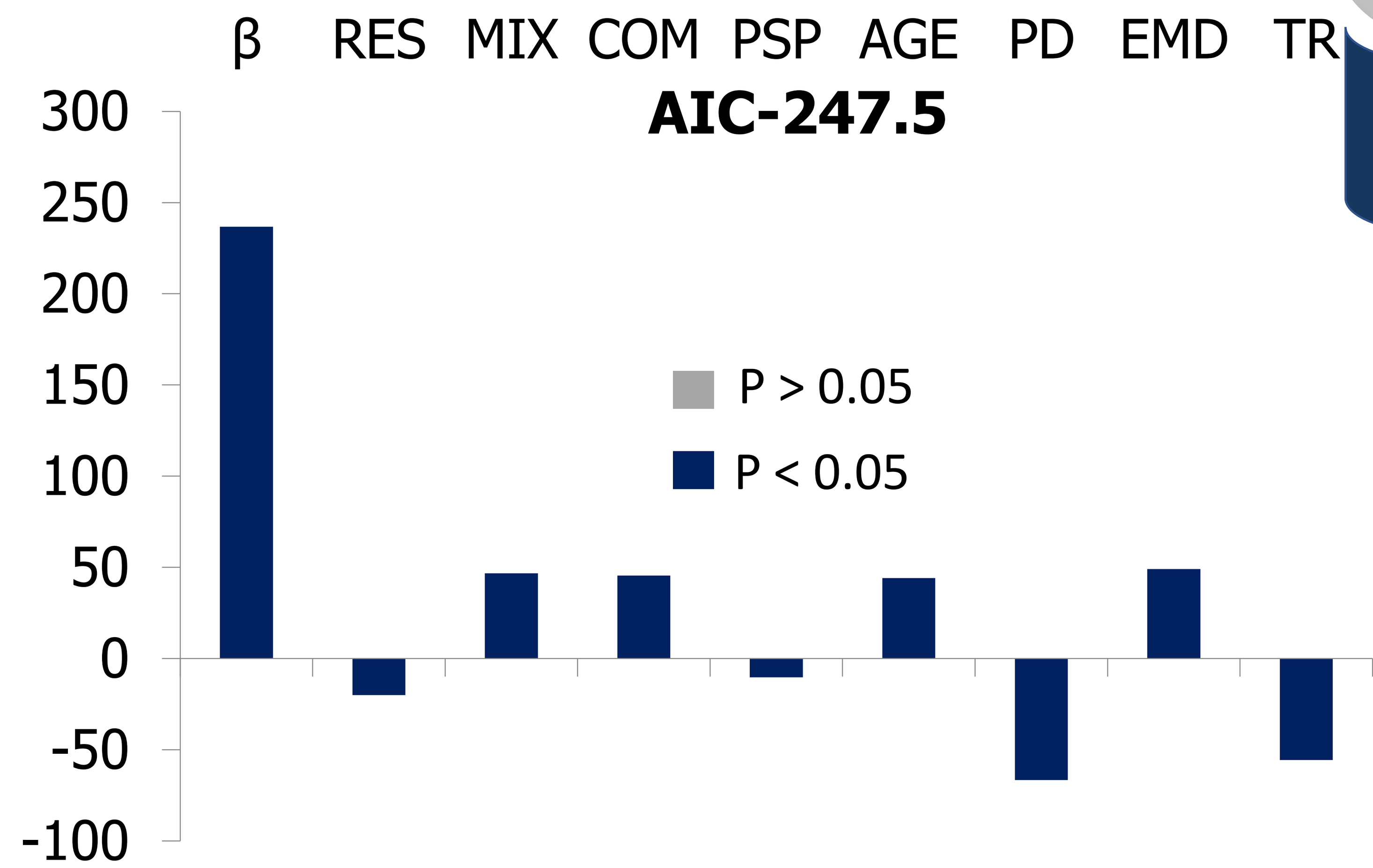
GAUSSIAN LINEAR MODEL ($R^2 = 0.93$)



NEGATIVE BINOMIAL MODEL ($R^2 = 0.92$)



LOG LINEAR MODEL ($R^2 = 0.92$)



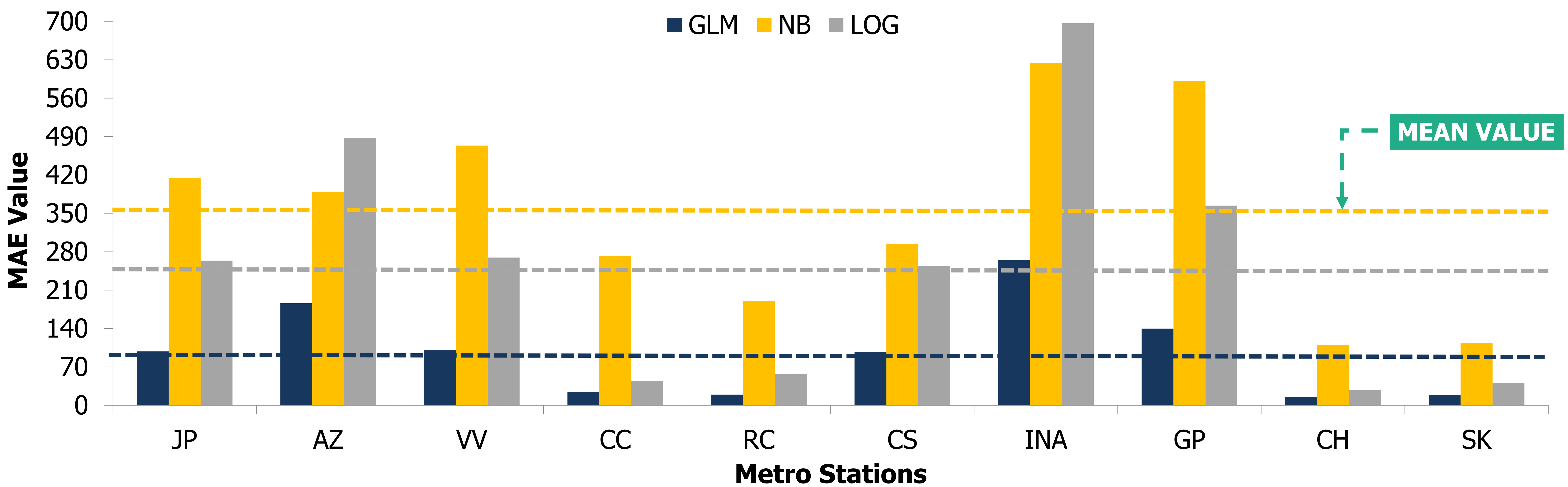
SUMMARY OF GAUSSIAN LINEAR MODEL

P value < 0.05	MIX, COM, AGE, EMD, TR	Coefficients Valid
P value > 0.05	RES, PSP, PD	Coefficients impact marginal
Positive Impact	MIX, COM, AGE, EMD	Will Increase the Dependent variable
Negative Impact	RES, PSP, PD ,TR	Will Decrease the Dependent variable

SUMMARY OF NB AND LOG MODEL

P value < 0.05	ALL VARIABLES	Coefficients Valid
P value > 0.05	--	Coefficients impact marginal
Positive Impact	MIX, COM, AGE, EMD	Will Increase the Dependent variable
Negative Impact	RES, PSP, PD , TR	Will Decrease the Dependent variable

COMPARISON OF ACTUAL PASSENGER DEMAND WITH PREDICTIVE VALUES OF ALL THREE MODELS



Regression equations of all three models

GLM	PF = [-8385.5 - 172.4(RES) + 489.8(MIX) +548.3(COM) - 211(PSP) + 12000(AGE) - 17.8(PD) + 42.3(EMD) - 5104.12(TR)]
NB	PF = [exp(8.629 - 0.0076(RES) + 0.0202(MIX) + 0.0304(COM) - 0.0105(PSP) + 0.6(AGE) - 0.0016(PD) + 0.00230(EMD) - 0.49(TR)]
LOG	PF = [exp(8.683 - 0.007(RES) + 0.019(MIX) + 0.03(COM) - 0.009(PSP) + 0.57(AGE) - 0.0016(PD) + 0.0024(EMD) - 0.5(TR)]

GLM model has the best R^2 value = 0.93, indicates a better fit of the model to the data & **lowest MAE** value, which means it has the best performance in terms of minimizing the difference between the predicted values and the actual values. Since R^2 value is high in all the models suggesting that the independent variables are successful in explaining the variation in the dependent variable. The **AIC** value is **144.87**, which indicates that **GLM model** has the best balance between goodness of fit and simplicity among the models.

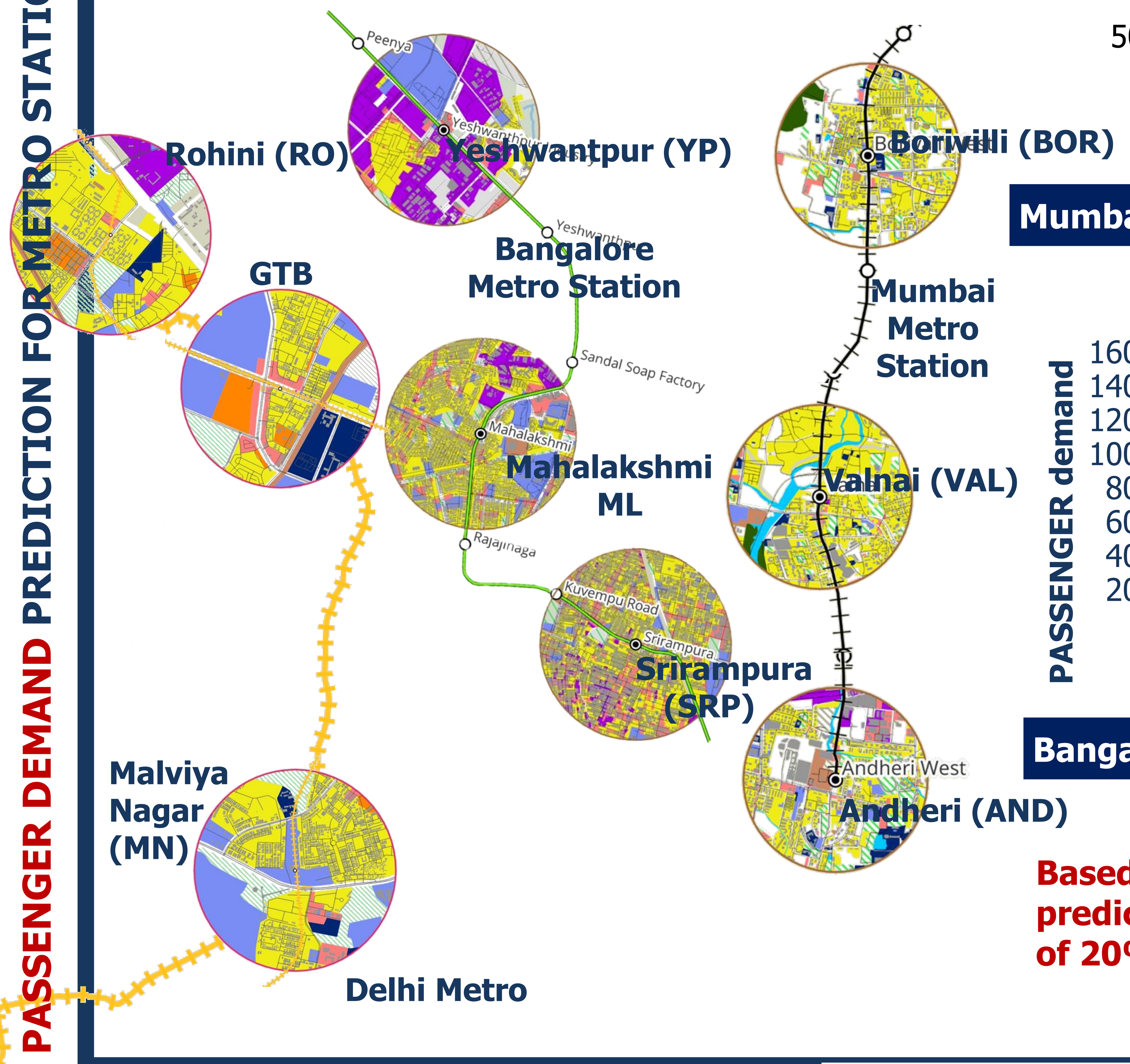
OBJECTIVE 3: EVALUATION AND COMPARISON OF ALL THREE MODELS

MODEL RESULTS AND VALIDATIONS

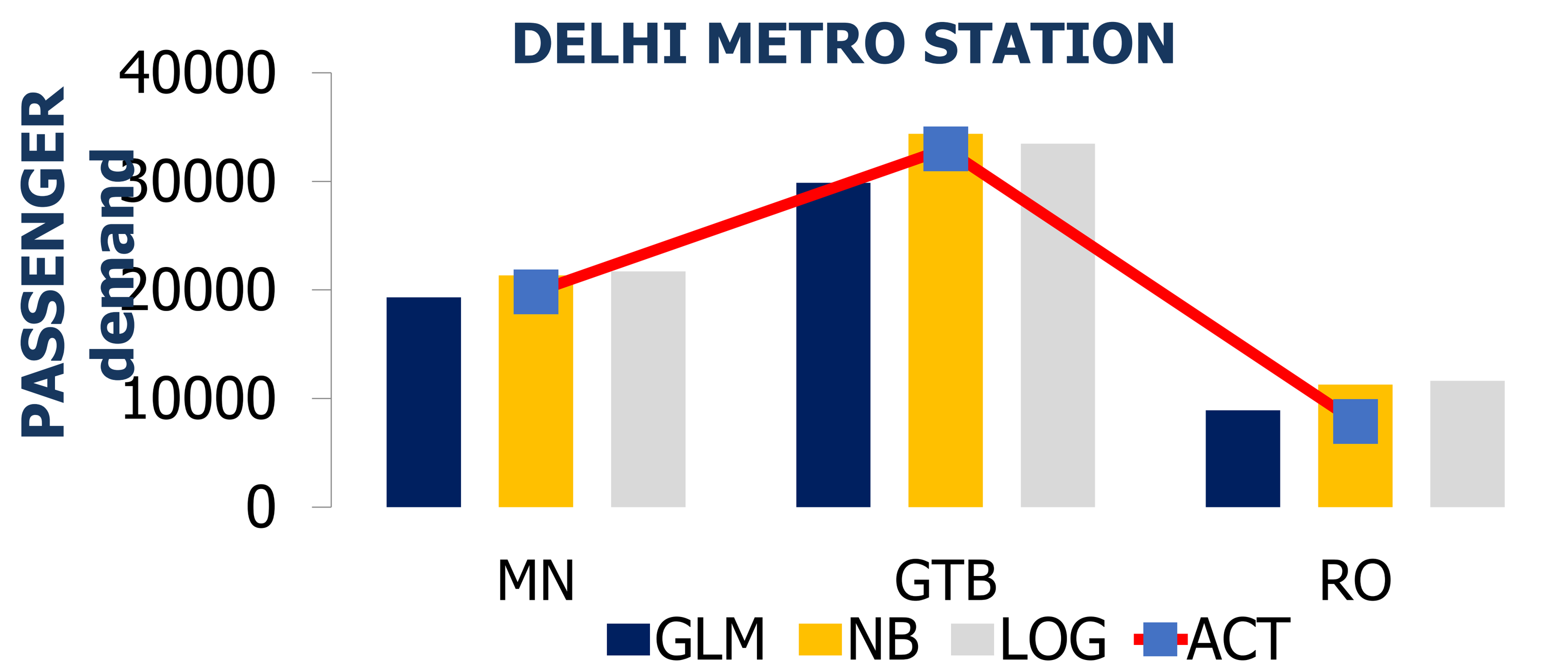
DATASET OF THREE METRO STATIONS OF DIFFERENT CITIES FOR VALIDATION

Station	RES	MIX	COM	PSP	PD	EMD	TR	AGE
D_MN	94.0	23.0	3.0	7.9	43.5	17.4	0	3
D_GTB	53.0	16.1	12.0	7.9	58.1	18.8	0	4
D_RO	84.0	37.0	17.0	4.3	50.1	18.6	0	2
M_BOR	55.6	0.0	2.5	2.1	4.6	4.6	0	2.5
M_VAL	41.8	0.2	1.2	1.0	6.4	6.4	0	2.5
M_AND	43.0	0.8	9.4	2.6	23.5	23.5	0	2.5
BG_YP	26.4	0.0	10.9	23.6	25.3	7.7	0	2.5
BG_ML	98.4	29.7	20.6	22.5	108.9	40.4	0	2.5
BG_SRP	107.1	39.7	32.8	12.7	283.3	200.2	0	2.5

LAND USE OF METRO STATIONS

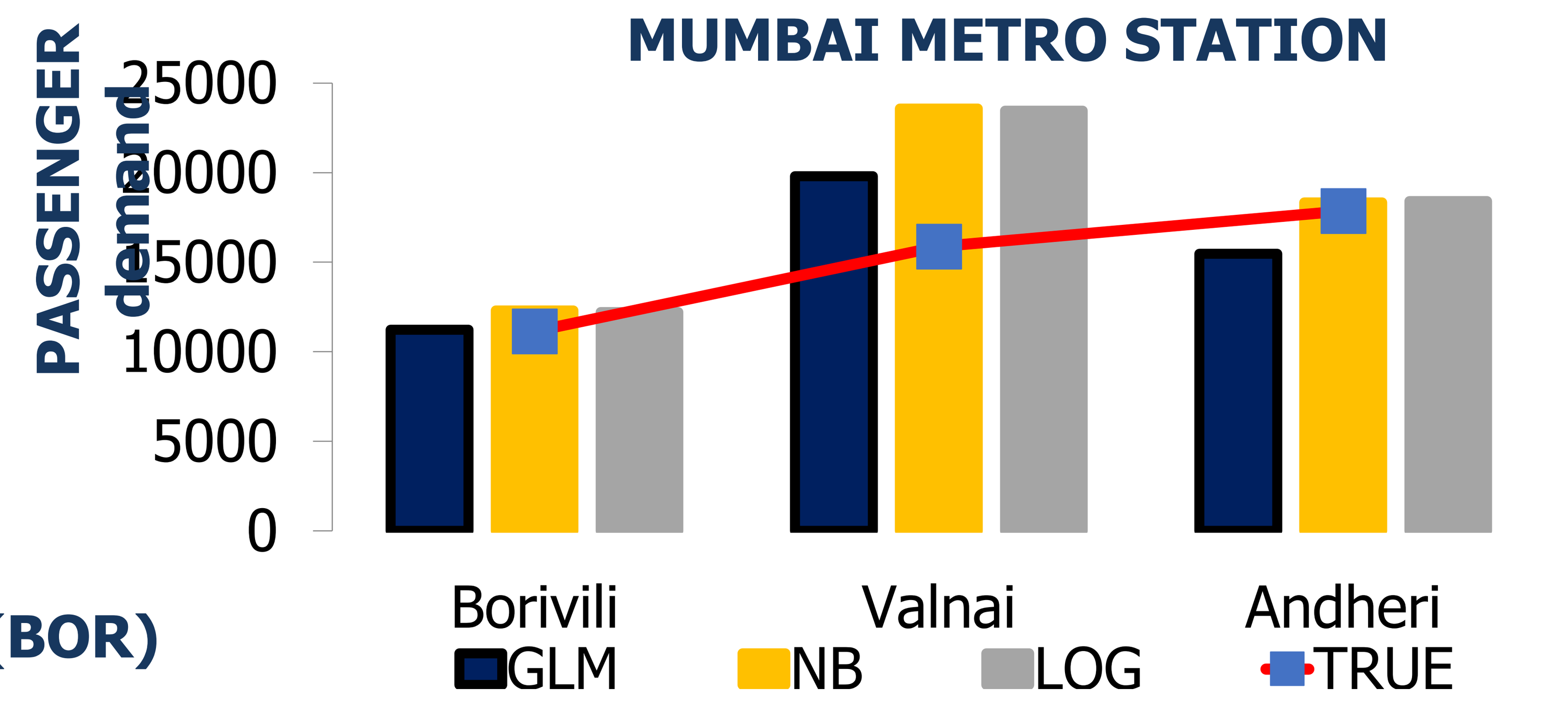


VALIDATION ANALYSIS OF ALL MODELS WITH PREDICTIVE VALUES OF OTHER THREE METRO



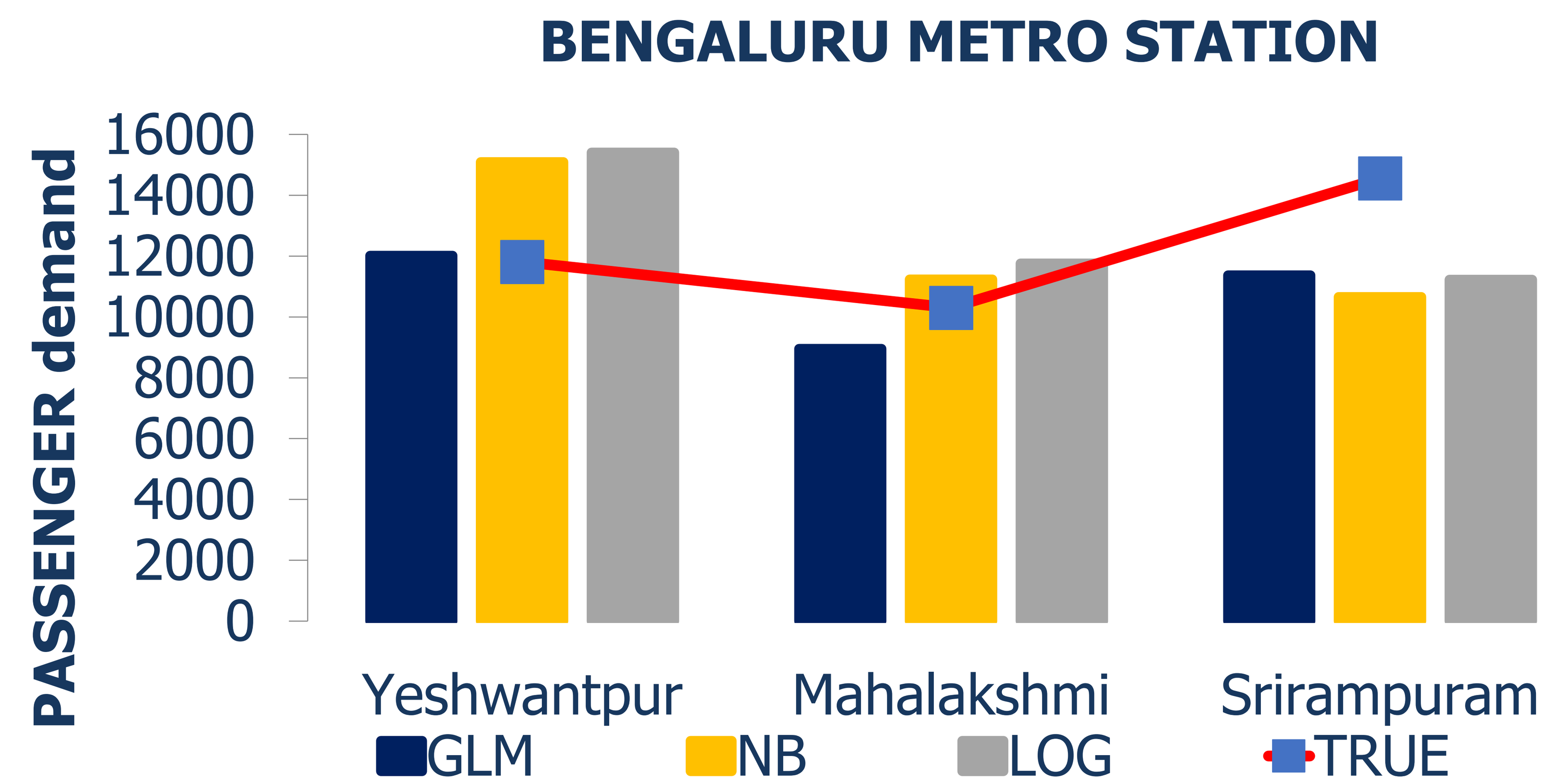
STATIONS	TRUE	GLM	NB	LOG
Malviya Nagar	19830	19325	21333	21702
GTB	32997	29879	34374	33458
Rohini	7872	8912	11284	11641
	MAE	**1554**	**2098**	**2034**
	MAPE	**8.4 %**	**18.4**	**19.6**

Delhi Metro Station (Line- Yellow Samaypur Badli-Huda City center)



STATIONS	TRUE	GLM	NB	LOG
Borivili	11141	11233	12303	12220
Valnai	15866	19796	23584	23460
Andheri	17858	15471	18338	18415
	MAE	**2136**	**3120**	**3077**
	MAPE	**13.0 %**	**20.6**	**20.2**

Mumbai Metro Station (Line- 2A Dahisar-Dahanukarwadi)



STATIONS	TRUE	GLM	NB	LOG
Yeshwantpur	11816	12020	15100	15408
Mahalakshmi	10302	8956	11237	11760
Srirampuram	14564	11372	10659	11229
	MAE	**1581**	**2708**	**2795**
	MAPE	**12.2 %**	**21.2**	**22.5**

Bangaluru Metro Station (Line- Green Nagasandara- Yellachenahalli)

Based on the MAPE and MAE values, the GLM model outperforms the other two models for predicting passenger demand at all stations. MAPE for GLM is 11% which is best, although MAPE of 20% and 21% of NB and LOG is good accuracy for regression model

OBJECTIVE 4: VALIDATION OF ALL MODELS IN OTHER CITIES

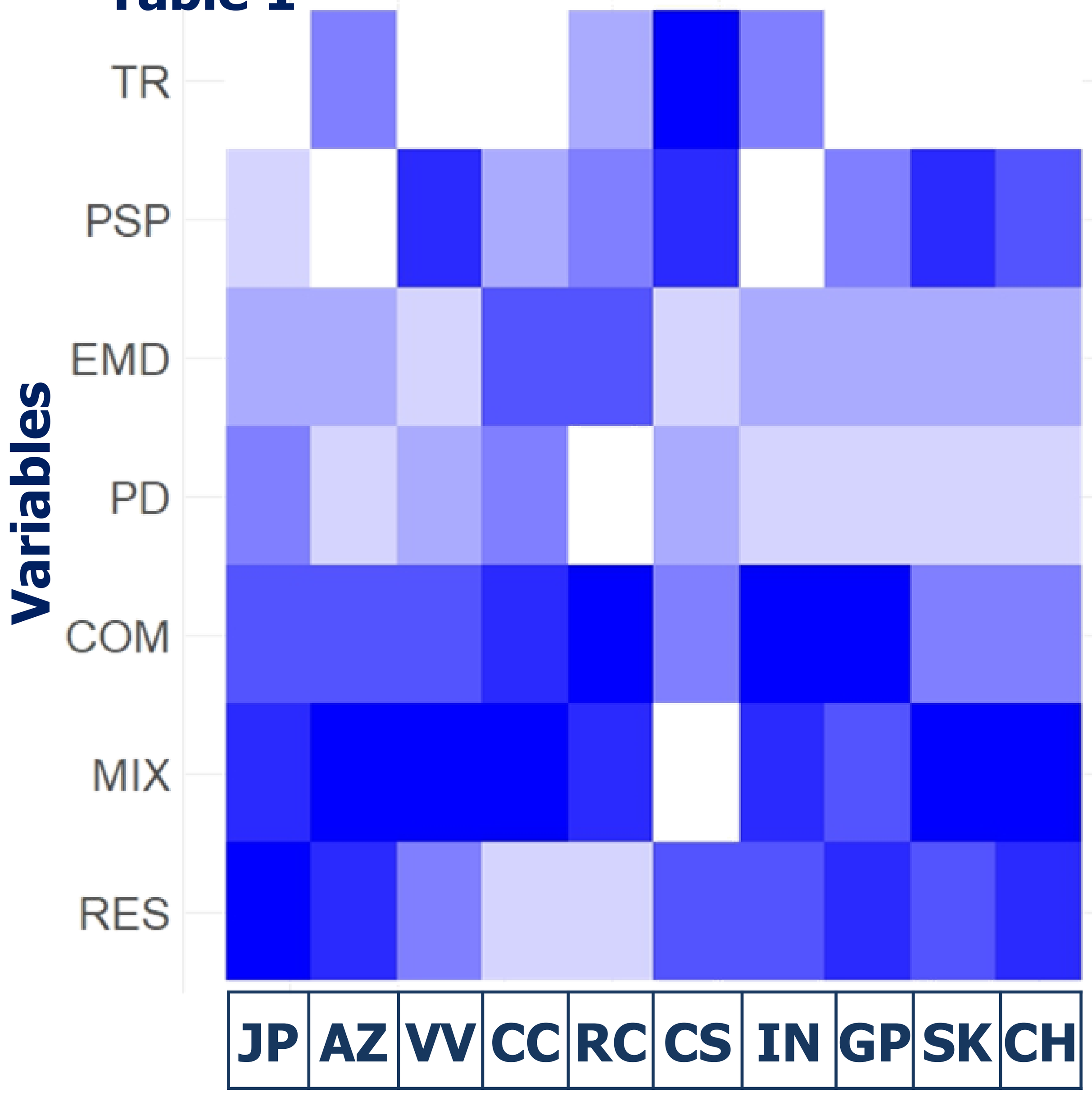
WAY FORWARD AND RECOMMENDATIONS

PASSENGER DEMAND PREDICTION FOR METRO STATIONS USING PROBABILISTIC MODEL

IMAPCT OF DEPENDENT VARIABLES ON EACH TRAINING STATION AND TEST STATION

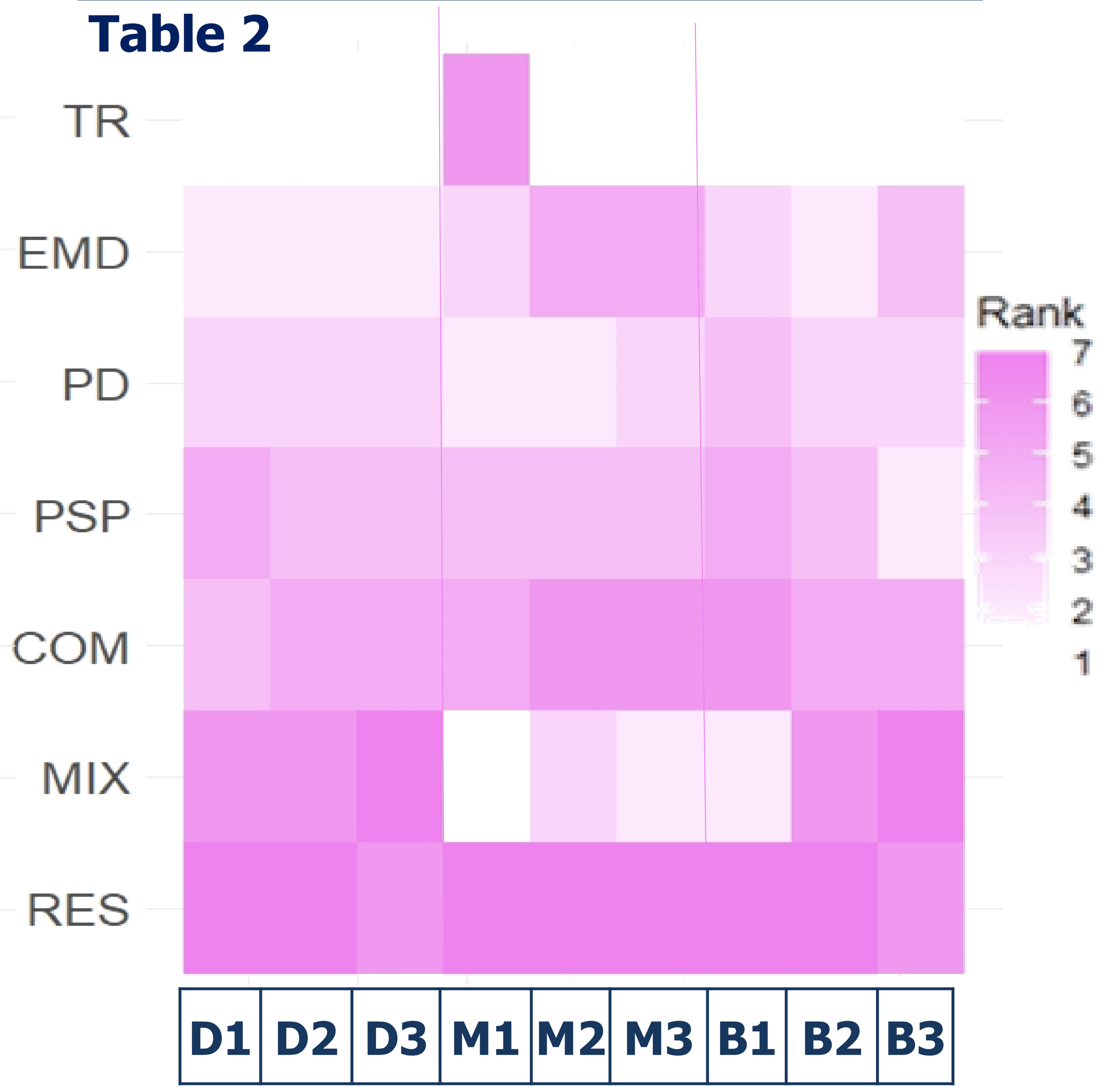
Heat map for case stations

Table 1



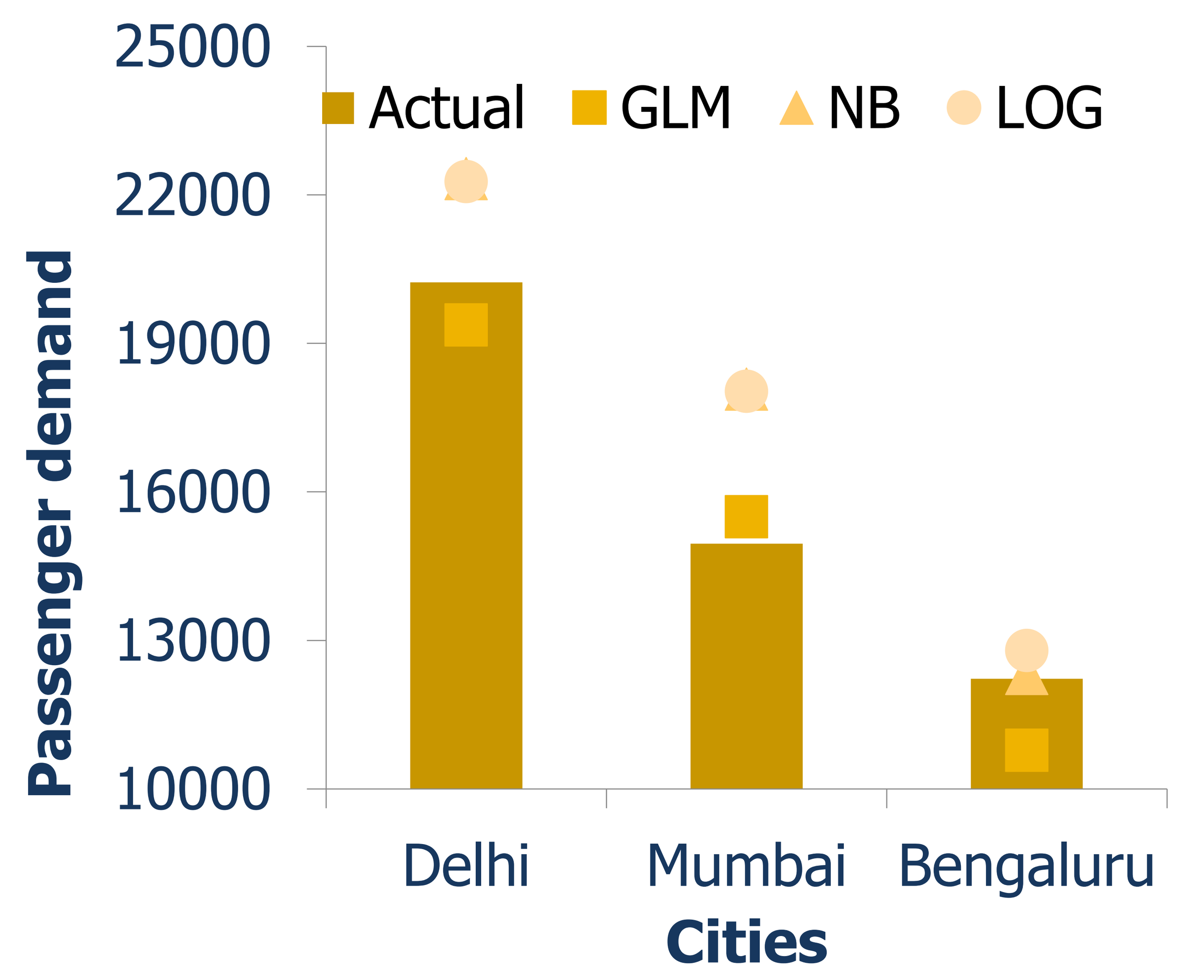
Heat map for validation stations

Table 2



Summary of overall impact of variables on different cities.

Case	Delhi	Mumbai	Bengaluru
MIX	RES	RES	RES
COM	MIX	COM	COM
RES	COM	EMD	MIX
PSP	PSP	PSP	PSP
EMD	PD	PD	PD
PD	EMD	MIX	EMD



Key Findings

Table 1 The most impacting variables are Commercial and Mixed land use. Table 2, Res and Com land use are the most impacting variable.

Residential is overall impacting variable.

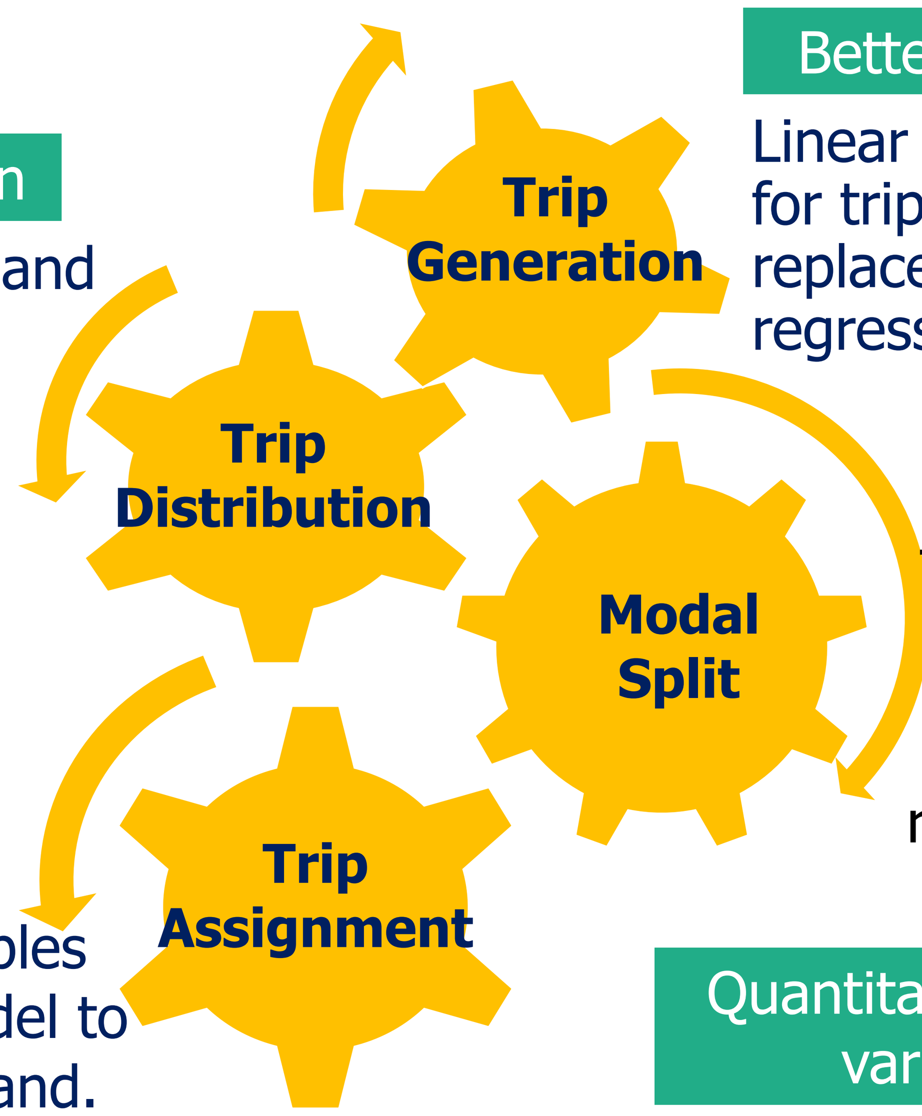
Gaussian Model has the best R^2 value, lowest AIC and best MAPE (11%) suggesting the model is better than the NB and LOG for predictive analysis of passenger demand.

RECOMMENDATIONS

1. Changes that can be made in the existing travel demand modelling in DPRs for better passenger demand prediction

Better Variable Selection

Zone attributes such as land use area and station characteristics can be included.



Better Accuracy

Linear regression technique used for trip generation can be replaced with probabilistic regression model

The estimated coefficient in the models provide insights into the strength and direction of the relationships between the predictors and demand.

Scenario Analysis

Different independent variables can be simulated in the model to assess their impact on demand.

Quantitative estimation between variables and demand

2. Usage of passages directly connected to metro stations for Residential areas

Zacharias et al. (2014) . Connecting Tokyo station with Yeshu Commercial center using Pedestrian Deck

Thus, overcoming the historical disconnect between the transportation facility and the surrounding environment

Since, RES and COM are most impacting variables similar decks can be made connecting nearby areas

Residential Area

Commercial Area

3. Need of short term Passenger Demand Estimation

With, urban expansion and dynamic changes in Land uses, short term prediction of demand evaluate the need for adjustments in plans and infrastructure according to changing demand

WAY FORWARD

- Exploring the temporal aspect of passenger demand by considering time-series data that can capture temporal variations and provide estimates which can help in improving operations
- To explore the use of machine learning methods for transit forecasting, such as neural networks or decision trees that may have better accuracy.

OBJECTIVE 4: RECOMMENDATIONS AND WAYFORWARD