Modelling Road Traffic Fatalities in India

Report on Fulfillment of Statistics-1 Project

Balaji Subramoniam P bmat1911
Pranav Krishna bmat1920
Rathindra Nath Karmakar bmat1923
Snehinh Sen bmat1935

***Abstract*—In this project, we try to reproduce and extend the analysis in Dr Rahul Goel's paper - "Modelling Road Traffic Fatalities in India". This report is a review of the methods we pursued and the sources we consulted to bring this project to fruition.**

I. Introduction

Being the second-most populous country in the world, India has a large number of road fatalities. According to the 2011 Census of India, pedestrians, cyclists and riders of two-wheelers modes contribute around 70 per cent of the commutes. Thus, it is unsurprising that most of the road fatalities occur for these categories of road users.

Until now, the Government of India has not undertaken steps to gather transport-related data in terms of traffic counts or travel surveys. As a result, most road injury models have been limited in their approach to account for the fact that there are multiple categories of road users with unequal risks. At most, vehicle registration data was used, which, among other things fails to account for the pedestrians and cyclists.

Considering the median wage in India, it is not surprising that the rate of vehicle ownership was around 6 per cent in 2011 and most commuters still walk, cycle or use Public Transport to reach their place of work.

Also, it must be noted that after the Economic Liberalisation of the country, from 1990 to 2015, the average year-on-year growth rate of 2W and cars was nearly 9 per cent. Leading to a doubling of the vehicular fleet every 8 years. The population has not had a similar increase and this denotes quite a significant mode shift from pedestrian/public modes to the private mode.

The motorisation of India is different from high-income countries in two marked ways

1) It is dominated by two-wheelers
2) Public transport includes non-standard modes

Firstly, for every car in India, there are as many as 5 2-wheelers. It is easy to see that keeping all other things constant, in the event of a crash, the rider of a motorbike is expected to be far more seriously injured than a person in a car.

Secondly, India's public transport system is, at best, patchy. The gap is filled by what the author called Intermediate Modes of Public Transport- autorickshaws and tuk-tuks, common in many south-Asian settings. While these tend to have a smaller engine capacity than a car, the ground reality is that they are usually overloaded. The significant presence of these modes presents necessitates a study of their impact on road safety and usability in general.

Lastly, the elephant in the room, the lax enforcement of Traffic rules, must also be addressed. It is common knowledge that the enforcement of Traffic Rules in India is sketchy. Most Tier-1 and Tier-2 cities have some enforcement in certain areas and there is some semblance of order on the National and State Highways. Other than that, one would be advised to be alert on the roads at all times because, in our experience, anything is possible.

The main objective of the paper-"Modelling Road Traffic Fatalities in India" was to develop an ecological model of road traffic fatalities with the states of India as the areal units while considering the effect of state-specific confounders.

A model of the below general form (commonly used in injury modelling) is used to realize the same. It is a log-linear relation written in exponential form.

$$n = M_1^{\alpha_1}.M_2^{\alpha_2}...M_n^{\alpha_n}.e^{\sum_{i=1}^{m}\beta_i.x_i}$$

where,

n denotes the annual road fatalities
M_i represents the travel distance or volume of road users corresponding to mode i.
α_i for i in $\mathbb{N}_n$ denotes the respective exponent of M_i
x_i for i in $\mathbb{N}_m$ represents a predictor variable controlling factors apart from volume and travel distance.
β_i is the coefficient corresponding to x_i

The developments in this paper pertain only to macro-level models viz. models which operate on large sections of the population like city wards, traffic zones and municipalities. This study is unprecedented for two reasons: Firstly, the dependent variable (the annual death count of all road users) is more general compared to the annual death count specific to one mode. Secondly, the model accounts for many modes of travel ($\geqslant 6$), consistent with the heterogeneous traffic patterns in India. This is in contrast to almost all[1] previous literature which consider only two modes with the conflicting mode as the dependent variable.

The motive to develop this model is twofold:

- Analytically, to derive conclusions about the correlations of the fatality rate with different modes from the estimated exponents obtained post modelling.
- Speculatively, to "predict" using the model the impact of futuristic changes in travel patterns on road traffic fatalities.

[1] [9] Elvik R (2016) modelled pedestrian injuries using the volume of cars and cyclists as dependent variables. All other preceding studies considered only two modes according to the author.

As a side note, the model is developed using commute distance opposed to the volume for two reasons: Due to inconsistent size of units of analysis and to account for changes in injuries in the case where two portions of the population use the same mode, but travel different distances.

In 2011, the Census of India introduced questions regarding the commute of workers. These questions were asked from a subset of all workers—the category called 'other workers'. This category excludes those involved in agricultural or household-based activities. The category of 'other workers' represents 42 per cent of all the workers in India. The two questions on commuting included mode of travel and one-way distance (in kilometres) from residence to place of work, and in the former one, only one mode could be selected, disregarding the multi-modal characteristics of most commutes using public transport. It is safe to assume that commuters answered with what one might call the main or most significant mode of travel. There are 9 options for the travel modes: (1) walk (2) cycle (3) moped/scooter/motorcycle (4) car (5) tempo/auto-rickshaw/taxi (6) bus (7) train (8) water transport (9) any other, and an option of 'No Travel'. Category 3 is referred to as motorised two-wheelers (2W), and category 5 as IPT.

One might go out on a limb and ignore the categories Water Transport, Other and No Travel as reported in the Census, because their effect on road fatalities is at best, undecidable. Also, these are less than 2 per cent in the Census data and can be ignored safely. For each mode, Census has reported mode-specific count of workers classified into 7 distance categories: 0–1 km, 2–5 km, 6–10 km, 11–20 km, 21–30 km, 31–50 km, and > 50 km. Walking has been re- ported up to 10 km, and cycling up to 30 km. The data has been reported only at the aggregate level of states and districts, with further classification into rural, urban and total. The total portion has been used for the analysis because the Government does not provide segregated data regarding urban and rural fatalities.

Also going out on a limb, we ignored the data for Lakshadweep and Andaman and Nicobar Islands, as their population is almost negligible. Also, it must be noted that the state-level makeup of India was different in 2011 compared to 2020. Specifically, the creation of Telangana and Ladakh.

As a side note, an inevitable limitation of estimating the commute distance using the census data is that it does not include road deaths from all road trips. The Ministry of Petroleum and Natural Gas conducted a study to estimate the modal petroleum and diesel consumption in the country[2], which is a good measure of the total distance travelled using different modes. The Pearson correlations of annual petroleum consumption at the state level for Two-wheelers and Cars with the respective commute distances were estimated to be 0.98 and 0.92 respectively.

For the Bus mode, the correlation between the state-wise total distance travelled reported by the respective State Road Transport Undertakings and the commute distance travelled by buses was estimated to be 0.89 .

[2]This study could not be accessed

The same could not be carried out for the other modes (Walk, Cycle and IPT) owing to unavailability of data. Nevertheless, taking into consideration the high correlations for the above three modes, we can assume similar patterns for the other modes and content ourselves with the commute distances obtained from the Census data.

The above-mentioned information is sufficient the purpose of this report[3]. Notwithstanding, for the sake of completeness, it is worth taking a look at some other related variables: Proportion of population living in urban areas, length of National Highways and population density. The Census classified the population into urban and rural categories; The state-wise length of National Highways is reported by the Ministry of Road Transport and Highways ([6] MoRTH 2011). The population density (in persons per sq. km) is calculated using the data in the web portal of the National Remote Sensing Centre ([8] NRSC) by dividing the population of each state by the sum of the respective urban and rural built-up areas (in sq. km).

II. Data and References

The following data were required for each state-

[1] Fatality Rate for road deaths

[2] Number of road deaths for each mode of travel

[3] Average Commute distance for each mode of travel

For item 1, we got the data from the yearly "Accidental Deaths and Suicides in India" report published by the National Crime Records Bureau.

For item 2, we got the data from the yearly "Basic Road Statistic of India" report published by the Ministry of Road, Transport and Highways.

For item 3, we modelled the Distance-Decay Functions of Work Trips in India and the mean was calculated analytically.

III. Formulae, Targets and Applications Used

A. A list of distributions and properties

The following distributions were used by the author to model Commute Distance for each mode in every unit. Each distribution was considered to be dependent upon two parameters.

- Lognormal
- Weibull
- Exponential

Additionally, unlike the original distributions which reach zero asymptotically, our data vanishes after a certain maximal distance. So, we introduce the concept of Truncated distributions.

Say $F(.|\alpha,\beta)$ is the CDF of some distribution L. Then we define $F_D(.|\alpha,\beta)$ of $F((.|\alpha,\beta,Max=D)$ as

$$F_D(x|\alpha,\beta)=\frac{F(x|\alpha,\beta)}{F(D|\alpha,\beta)}\mathbf{1}[x<D]+\mathbf{1}[x\geq D]$$

Note that Right Continuity, Limiting values and Increasing nature of F are preserved for F_D. So, it is still a CDF. We say

[3]The following discussion is relevant (only) to Regression models 2, 3 and 4 mentioned in the succeeding sections

$F_D(.)$ is the CDF of the distribution L right truncated at D. Also, the expectation and variance we shall speak about are conditioned to the constraint that $x < D$. Some were derived analytically, while some were recounted and stated from prior sources.

The Computation was done in R and VBA (Codes Provided Later). We provide a brief overview of each distribution.

§Lognormal

A well-known distribution, the author has used the 2-Parameter Lognormal Distribution with the following properties.

- CDF:

$$F(x|\alpha,\beta) = 0.5 + 0.5 * erf((\log(x) - \alpha)/(\sqrt{2}\beta))$$

- Truncated CDF: (as seen before)
- Mean (Truncated) :

$$\frac{e^{\alpha+0.5\beta^2}\Phi(A - beta)}{F(D|\alpha,\beta)}$$

- Standard Deviation (Truncated) :

$$\sqrt{\frac{e^{2(\alpha+\beta^2)}\Phi(A - 2beta)}{F(D|\alpha,\beta)}}$$

Here, $A = \frac{\log(D)-\alpha}{\beta}$ and Φ is the CDF of a standard Normal $\mathcal{N}(0,1)$

§Weibull

The following are the properties of the two-parameter truncated Weibull Distribution.

- CDF :

$$F(x|\alpha,\beta) = 1 - e^{-(x/beta)^{alpha}}$$

- Truncated CDF : (as seen before)
- Mean (Truncated) :

$$I\gamma((D/\beta)^{\alpha}, \frac{1}{\alpha} + 1)$$

- Standard Deviation (Truncated) :

$$\sqrt{\beta I\gamma((D/\beta)^{\alpha}, \frac{2}{\alpha} + 1) - (I\gamma((D/\beta)^{\alpha}, \frac{1}{\alpha} + 1))^2}$$

Here, $I = \frac{\beta}{F(D|\alpha,\beta)}$ and γ is the lower incomplete gamma function defined as.

$$\gamma(x,a) = \int_0^x t^{a-1}e^{-t}dt$$

§Exponential

The following are the properties of the two-parameter truncated Exponential Distribution.

- CDF :

$$F(x|\alpha,\beta) = 1 - \beta e^{-(\alpha x)}$$

- Truncated CDF : (as seen before)
- Mean (Truncated) :

$$\frac{1 - (\alpha D + 1)e^{-(\alpha D)}\beta}{F(D|\alpha,\beta)\alpha}$$

- Standard Deviation (Truncated) :

$$\sqrt{\left(\frac{-\beta e^{-\alpha D}}{\alpha^2 F(D|\alpha,\beta)}\right) t(\alpha, \mu - D) + t(\alpha,\mu)}$$

Here, μ truncated mean of this distribution given α, β, D. Also, $t(x,y) = xy(xy-2) + 2$

B. A look into the regression model

A Poisson regression model[4](mentioned below) is used to model the annual fatality count with the mode-wise commute distances as predictors using Bayesian Methods. R-INLA (Integrated Nested Laplace Approximations) is used to fit the model.

$$y_n = Poisson(f_n)$$
$$log(f_n) = log(e_n) + \beta_0 + \beta X_n + \delta_n$$
$$\delta_n \sim N(0, 1/\tau_n)$$
$$log(\tau_n) \sim logGamma(1, 0.0005)$$

y_n : Observed annual fatality count of road users in state n
f_n : Expected annual fatality count of road users in state n
β_0 : Intercept
e_n : Exposure (Population of state n)
X_n : Vector of explanatory variables(natural log of mode-wise commute distance[5])
β : Vector of fixed effect parameters.
δ_n : Uncorrelated Heterogeneity/Unstructured Error
τ_δ : Precision of the distribution of δ_n

Details on the calculation of X_n:

For the total population of the n^{th} state, $(X_n)_i$ is calculated as

$ln(n_{tot,obs,i,s} \times E[X_{mod,i,s}])$

where $n_{tot,obs,i,s}$ is the total number of people of state s (n^{th} state) who reported the i^{th} mode as main mode of travel;

$X_{mod,i,s}$ is the random variable having the modelled distribution for the distance bins of i^{th} mode and state s.

Then X_n is the vector whose i^{th} component is $(X_n)_i$.

For those cases where the main mode of travel is Public Transport of some kind, the walking distance is underestimated since the Census data deals only with the main mode of travel. We account for this in this model by assuming 1 km

[4]As discussed in the meeting, in this report we shall replicate only "Model 1" made by the author

[5]The commute distance corresponding to a particular mode M and state S is defined as the product of the mean distance travelled by people in S using M and the number of people in S using M

of walking distance corresponding to each trip longer than 1 km made using Public Transport(Includes Bus, IPT and Train)

C. List of Applications Used and The Functions

The following programming languages/softwares were used for carrying out computations. The functions referred to have been mentioned as well (this does not include the standard ones like sqrt, exp, etc. and our functions):

- R-Studio: The R programming language is used for statistical computing and graphics. R-Studio is an IDE (Integrated Development Environment) for the R language.
  - kmeans(x, centers, ...): To perform k-means clustering on a data matrix. We have used the first two parameters, x- which stores the matrix and centers- a vector of initital cluster centers
  - png(filename = "Rplotwidth = 480, height = 480, units = "px", ...): To output the argument image file in a specific format and adjust its properties
  - pie(x, labels = names(x), edges = 200, radius = 0.8,..): To draw a piechart from the argument vector
  - barplot(height, width = 1, ...): To draw a barplot for the argument data
  - erf(z): The Gaussian error function
  $$\frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$$
  - gamma(x): The Gamma function
  $$\int_0^\infty t^{x-1} e^{-t} dt$$
  - incgam(x,a): The Upper Incomplete Gamma function
  $$\int_x^\infty e^{-t} t^{a-1} dt$$
  - inla(formula, data, family, offset, ...): inla performs a full Bayesian analysis of additive models using Integrated Nested Laplace approximation

  The erf, gamma and incgam function were available under package "pracma", which contains advanced functions for numerical analysis
- Microsoft Excel: Spreadsheet software developed by Microsoft. We mainly used this in conjunction with Visual Basic For Applications
- Visual Basic for Applications: An implementation of Microsoft's event-driven programming language Visual Basic. We used the version that comes with Excel 2016 and later. Most functions we used are either the standard ones or defined by use. Besides, we had
  - erf: Defined in the same way as before
  - Solver functions- The GRG (Generalised Reduced Gradient) non-linear solution method for Excel's Solver function was implemented using Visual Basic's Solver controls. The method starts with an initial point, performs a gradient search to determine the direction of cost decrease and varies the function to reduce the cost. This process is continued till it reaches a local optimum.
    - SolverReset: Resets the Solver
    - SolverAdd: Adds Cell references for Constraints, variable cells, objective cells, etc
    - SolverOk: To specify that the optimization problem has been defined
    - SolverOptions: To specify other solver options like precision, CPU time, etc
    - SolverSolve: To start the optimization procedure

IV. The Processes

A. Calculation of parameters and best fitting distribution using GRG and VBA

The method of determining the distribution is detailed in the next section. Here, we look at the method of finding the parameters of the distribution. The objective is to find the best fitting distribution, one which minimises the difference between the observed and expected values of distance counts. For this, we optimise the chi-squared statistic given by

$$\chi^2 = \sum_i \frac{(n_{i,obs,m,s} - n_{i,mod,m,s})^2}{n_{i,mod,m,s}}$$

where $n_{i,obs,m,s}$ represents the observed value of the distance count in the i^{th} bin for mode m and state s; $n_{i,mod,m,s}$ represents the expected value for the same, given by $n_{total,obs,m,s} * \mathbb{P}(i)$ where $n_{total,obs,m,s}$ represents the total observed count for mode m and state s, $\mathbb{P}(i)$ represents the probability that the modelled distribution takes values in the i^{th} distance bin (calculated from the cdf) The database was loaded into MS Excel; the cdfs of the truncated exponential, lognormal and weibull distributions were entered into Visual Basic, along with the respective functions for $n_{i,obs,m,s}$, $n_{i,mod,m,s}$, $n_{total,obs,m,s}$ and χ^2. Excel's GRG solver method was implemented by defining a "solver" method which executes the solver :

- separate cells were allocated for the two parameters α and β. The initial values were set to 1 for both.
- The GRG solver was programmed to minimise the χ^2 function by changing the values of α and β, subject to the constraints that both are upper-bounded by 20. In a few cases, the upper bound was set to 65 for both
- When the solver caused an error due to division by 0, the initial condition was set to a value close to 1 and the calculations were repeated.

The solver solution gave values for α and β for which χ^2 attains a local minimum.

B. Deciding Distribution using Pearson Correlation

Based on previous models for distance decay functions, Exponential, Lognormal and Weibull distributions were considered for modelling the current distance data. The two-parameter exponential distribution was preferred over the others for walking, as the likelihood is highest for trips close to zero and drops sharply thereafter. The minimisation procedure

State	Mode	Distribution	D	alpha	beta	Mean	Standard Deviation
BIHAR	Cycle	Lognormal	30	1.70	1.45	6.96	9.80
ANDHRA P	Car	Weibull	200	0.82	15.35	17.07	20.773
ODISHA	Walk	Exponential	10	0.21	0.69	2.22	2.31
GOA	Train	Weibull	200	0.76	20	22.89	28.78
MEGH.	IPT	Lognormal	100	1.41	1.53	9.56	17.48

TABLE I: Table of Mean and SD of State/Modes

outlined in the previous section yielded a value of the chi-squared statistic of the order of 10^{-10} and smaller, showing that it provided a perfect fit for walking.

For all other modes, except cycling, the lognormal distribution was used for an initial analysis and chi-squared values, modelled counts for each distance bin was calculated. Thereafter, the Pearson correlation coefficient between the vectors of modelled and observed counts was calculated.

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \times \sqrt{\sum_i (y_i - \overline{y})^2}}$$

where $(x_1, x_2, \dots x_n)$ and $(y_1, y_2, \dots y_n)$ represent the vector of observed and modelled counts for distance bins respectively. A correlation of 0.99 or more signified a good fit and the lognormal model was retained for the respective entries. For others, the Weibull distribution was used for modelling.

As for cycling, some discrepancy was observed in the data: barring the NCT of Delhi, it was observed that the count for the 20-30 km bin was larger than or equal to the count for the 10-20 km bin. The histogram of any modelled distribution is likely to have a negative slope over the $> 10km$ range as the frequency of cycling reduces over such large distances. To deal with this discrepancy, the modelling was done after combining the two distance bins into one for cycling and the lognormal distribution was used. Thereafter, the modelled distance count for the two bins was calculated separately and it was observed that the lognormal distribution provided a good fit; in particular, the larger counts for 20-30 km shifted to 10-20 km bin.

C. Means, Standard Dev and Mode Shares

After the parameters were calculated and the distributions were verified, we use R to calculate the mean and Standard Deviation.

The meanvar.r file was used in which our table with optimal alpha-beta and best fit distribution was drawn. Using that we use the functions `meannew(dist, alpha, beta, D)` and `stdnew((dist, alpha, beta, D)` were used to calculate mean and standard deviation.

Here, the function reads `dist` as the distribution in Capitalized Format `('Weibull', 'Lognormal','Exponential')`. `alpha,beta` are the respective parameters as in 3A and `D` is the maximum distance.

One might doubt whether or not the maximum distance for the seemingly unbounded ones is feasible. In any case, A Deviation by 25% on either side of the distribution does not affect the mean by more than 0.5 (Which is pretty less in aforementioned situations). Thus, from a sensitivity point of view, this verifies that it is indeed proper enough to consider restricting the support as given.

This way, after computing the mean and standard deviation of the modelled data, we save them in the 27th and 28th Column of `WorkingOutOurDataMeanSD(AfterMean).csv`.

A summary table of means and Standard Deviation along with alpha, beta, D and Distribution for 5 Sample States and Modes are given in Table I.

After this calculation was done, we create three new files, namely

- ModesNumberShare.csv - To Write How The Total Number of participants are shared based on modes in each unit
- ModesAvgAuthor.csv - To Reformat Computed Means of author for each of the states and save it in a format similar to `ModesNumberShare.csv`.
- ModesAvgUs.csv - Similar to `ModesAvgAuthor.csv` but with means computed by us.

These Tables Shall be helpful in computing Modes Distance Share (Fraction of Total Distance Travelled for each mode in a fixed state unit) and hence fruitful for Cluster Analysis. (Check next section)

After creating these files we create mode share graphs for each table using the code present in `graphsmodeshare.r`. The Graphs of our Interest our the standard Pie Chart Representing Share of total number of people considered in each mode and a detailed Stacked Bar chart of distance bin wise as well as total mode share. Some examples are given in the outputs section.

D. Mode Distance Shares and Cluster Analysis

As mentioned earlier, we compute Mode Distance Shares and use it to split our data into clusters. Before we go into the data and process, let us first look at what Cluster Analysis is.

Cluster Analysis is the process of dividing our data of interest into clusters having a certain common property or exhibiting closeness to themselves or certain fixed points. There are many cluster analysis techniques. For our purpose, the most suitable has been deemed to be a K-means cluster analysis.

In this, suppose we have n datapoints (generally vectors). Our sole target is to create K clusters such that if we take the sum of squares distance of each element from the mean of the cluster in which it belongs, this is minimized. The target to be minimised is called the Within Cluster Sum of Squares.

Mathematically, our target is to do the following

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

where $(\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_n})$ is the vector of n quantities and S is one of all possible clustering or partition of these n objects into k clusters, say $S_1, \dots, S_i$. once S is selected, $\mu_i = \text{mean}(S_i)$

There is a very beautiful resemblance of this with the Sum of Squares Error in Anova test. There might be multiple minima and up to a range of possible error, our target clusters

might differ (Check Section VII) on each run (let alone wrt the Author).

However, the clusters formed are generally acceptable, and we have presented our results accordingly.

One more interesting question is to decide what is k? Or, framed otherwise, how many clusters?

Note that as we increase the number of clusters, the minima obtained will decrease further and further. Here, we follow our gut and authors prediction. It is observed that around 5, a visible knee appears in the graph of least $WCSS$ and the number of clusters. No doubt this might keep on changing with each run (as computer looks for local minima) but after several runs, this was observed to be more or less consistent. By a knee, we refer to the fact that after that value, we observe a flattening of the curve and a relative drop in the absolute slope of the graph (line plot). This is why we end up doing it for 5 clusters.

Now, coming to the code. As stated earlier, we do cluster analysis on the modes distance share. To optimise database, we do both of them in the same step using the function `createcluster` created in `ClusterAnalysis.R`. This function first creates a dummy table (Actually, the mode distance share table) and does a cluster analysis on the numeric entries to arrive at 5 clusters. The Cluster Analysis (basically an optimisation) is done using the in-built `kmeans` function, described before. After the clusters are found, we go to output table (`"AuthorCluster.csv"` or `"OurCluster.csv"`) in excel and sort it. Then we compute the mean of each column and using it as the central number, create a heat map. We save the same as (`"AuthorClusterHM.xlsx"` and `"OurClusterHM.xlsx"`) The results are shown and explained in the following sections.

The Algorithm used is the *Hartigan-Wong Algorithm* which works as follows :

1) The method is a local search that iteratively attempts to relocate a sample into a different cluster as long as this process improves the objective function.
2) Let $\varphi(S_j)$ be the individual cost of S_j defined by $\sum_{x \in S_j} (x - \mu_j)^2$, with μ_j the center of the cluster.
3) **Initial Step :** We allot data into K random clusters, say $\{S_j\}_{j \in \{1, \cdots k\}}$.
4) Next it determines the $n, m \in \{1, \ldots, k\}$ and $x \in S_n$ for which the following function reaches a maximum
$$\Delta(m, n, x) = \varphi(S_n) + \varphi(S_m) - \varphi(S_n \setminus \{x\}) - \varphi(S_m \cup \{x\}).$$
5) For the x, n, m that reach this minimum, x moves from the cluster S_n to S_m
6) The algorithm terminates once $\Delta(m, n, x)$ is larger than zero for all x, n, m.

For ease and efficiency of calculation, one can rewrite Δ as

$$\Delta(x, n, m) = \frac{\mid S_n \mid}{\mid S_n \mid -1} \cdot \|\mu_n - x\|^2 - \frac{\mid S_m \mid}{\mid S_m \mid +1} \cdot \|\mu_m - x\|^2.$$

Akin to the local analysis and optimisation used for determination of parameter of distributions, we arrive at a "Local Minima". However, finiteness of the number of clusters and large datasets give us hope that the output is apposite. I'm not sure that apposite is the right word here. It is similar to apt, and we want our classifications to reflect some pattern/order in the real world. Feel free to ignore.

Some other Algorithms for K-means Clustering are :

- Lloyd: Iterative recalculation of means after reassignment to closest mean in each step - might not terminate, but if it does, the output is a Global Minima. Centroids are recalculated after all reallocations.
- Forgy: The same as Lloyd, but ith an underlying assumption of continuous distribution. (Though R discretises both of them)
  what is ith in the previous para?
- MacQueen: Similar to Lloyd as well as Hartigan Wong - Based on recalculation/reassignment of the centroid, but at the same time updates are made whenever one datapoint changes cluster. So at the cost of minimal memory and computation, the output is generally increased, making output Global and keeping the process pretty smart.

Though we thought that K-means clustering is the best, one can choose other methods as well, some of them are :

- Partition Clustering (Similar to K-means
- Hierarchical Clustering
  - Agglomerative
  - Divisive
- Model-Based Clustering - effective for those with an a priori structure and knowledge. For example, using a regression model to categorise.
- Fuzzy Clustering Algorithms
- Density-Based Clustering - Good for large data sets and very much used in unsupervised machine learning
- Neural and Other Probabilistic Clustering Methods.

E. Regression and Sensitivity Analysis

Four models (Model 1, Model 2, Model 3, Model 4 based on the above-mentioned structure were analyzed with each succeeding model accounting for more explanatory variables. The explanatory variables accounted for by each model are mentioned in the table below. For the sake of brevity, in the table, we mention only the additional explanatory variable(s) used, compared to the previous model.

Model	**Components of the vector of Explanatory Variables (X_n)**
Model 1	The logarithm of the Commute Distance corresponding to each mode (Bus, IPT, Car, Walk, Bicycle, Two wheelers)
Model 2	Additionally the logarithm of diesel consumption
Model 3	Additionally the length of National Highways
Model 4	Additionally the proportion of urban population and the population density

Since we shall be concerned only about Model 1 in this report, we define X_n as follows:

$$X_n := (ln(bus), ln(auto), ln(car), ln(walk), ln(bicycle), ln(X2w))$$

where, $bus, auto, car, walk, bicycle, X2w$ refer to the total commute distance travelled by people in the state indexed n, using Bus, IPT, Car, Walk, Bicycle and Two wheeler modes respectively as their main modes of transport. vsoace2mm

We further make the natural assumption that the components of X_n are independent.

The data from Census 2011 and the National Crime Records Bureau(2010-12) (for fatality rates) summarised in `"pptauth.csv"` were used and the model was fitted with the help of R, using the *R-INLA*[6] package to obtain the posterior distributions of the fixed effect parameters. Further remarks on the model and details on the implementation are annotated in the R codes.

To account for the underestimation of the walking distance, we already accounted in the regression model, 1 km of walking distance for every Public Transport trip longer than 1km. We carried out the sensitivity analysis by varying the walking distance assumption made in the regression model. We studied and calculated the posterior distributions for two cases where we assumed:

- No walking distance corresponding to each PT trip
- 1.5 km of walking distance corresponding to each PT trip with distance more than 1 km

The R codes `"model1.R"`, `"model1sens0.R"` and `"model1sens15.R"` were used for the regression modeling and the two cases of sensitivity analysis respectively.

F. Mode Shift

From the regression model results (which will be discussed later), we have the following equation.

$$f_n = k.bus^{0.09}.car^{0.063}.walk^{-0.250}.cycle^{-0.211}.X2w^{0.447}.auto^{-0.098}$$

where, k is a positive constant.

We studied the scenarios where road users gradually shift from their current mode of transport to a new mode of transport as they acquire new vehicles to observe the variation of annual road fatality rate with different amounts of shift, given the regression model. - We considered three natural shift cases:

- Shift from Two-wheelers to Car
- Shift from Walk to Two wheeler
- Shift from Cycle to Two wheeler

We obtained the average commute distance and mode share of each mode for each of the clusters[7]. For an illustration of the procedure, let us assume that all explanatory statements made in the next two paragraphs pertain to some specific cluster. The statements apply likewise for all clusters.

We define the baseline as the point where each variable in the above equation corresponds to the average modal commute distance of the cluster. Starting from the baseline, we proceed in steps by successively subtracting 0.5% of points of the mode share from the origin and adding it to the destination. The total commute distance is invariant in this process. We perform 9 steps if allowed[8]. We refer to the configuration of mode shares after a number(0-9) of steps as points. Hence, for a cluster, we obtain 10 points.

The Relative Risk corresponding to a point is defined as the ratio of the value of f_n calculated at that point to the value of f_n calculated at the baseline.

The points are obtained and the respective Relative Risks are computed for all clusters for the three mode shift cases. We plot the Relative Risk against the points in the increasing order of mode shift for all clusters, separately for the three shift cases.

V. Outputs, Explanations and Results

A. A summary table of means, standard deviation, alpha, beta.

In this section, we do a head to head comparison of means and standard deviations calculated by us with those reported by the author.

There are some deviations. However, upon verification, we realised that our values provide a better fit. Also, there were some discrepancies in the Author's data (discussed later).

To conclude, we give tables for different modes in Jharkhand to show these values, separately for the Author and Us. Throughout the paper, we shall analyze both of these data wherever possible.

The difference in the β values stands out. Upon closer analysis, we realised that this might be because of an entry error in the part of the author. IN the author's data for Weibull, it has always been recorded that alpha and beta are equal. Clearly, this is not at all a feasible assumption, as using those values not only was the chi-squared unoptimised but it wasn't even giving us mean and sd as expected. (Even using third party calculators). The same is discussed once again later on.

B. Mode Share (Just Number) Pie Charts and Bar Charts

In this section, we look at some graphs of mode share and try to explain why the mode share is as such in some example states. As a sample, we take and study Assam and Haryana.

§Assam

As stated earlier, we look at the total mode share (quantity) pie chart and the total distance bin wise mode share analysis. Along with them, few descriptive data tables are also provided.

[6]INLA (Integrated Nested Laplace Approximations) unlike MCMC (Markov Chain Monte Carlo) methods, is an approximation method. However, it is fast, easy to use and works well with complex models

[7]Please refer to the subsection 'Mode distance share and cluster output' under the section 'Outputs, Explanations and Results' for statistics and information on the clusters used

[8]It might not be possible(since mode share must be positive) in all divisions of states into clusters since the mode share of a concerned mode of transport might be very low in a particular cluster as was the case in the Author's division. In such cases, we perform the maximum number of steps allowed or 9 steps, whichever is minimum. However, in our division, it was possible for 9 steps to be performed for all clusters and all cases.

Mode	Total	Distribution	Maximum Distance	Alpha	Beta	Mean	SD
Cycle	949583	Lognormal	30	1.60	1.21	6.74	9.31
Bus	168791	Weibull	100	0.83	20	19.56	20.46
Car	73165	Lognormal	200	1.79	1.39	13.76	26.11
2W	553501	Lognormal	100	1.36	1.19	7.45	12.84
Walk	1259203	Exponential	10	0.24	0.71	2.18	2.24
IPT	174579	Lognormal	100	1.76	1.01	9.36	14.30
Train	112869	Weibull	200	0.53	20	25.51	37.59

TABLE II: Data Description for Jharkhand: Using Our Data

Mode	Total	Distribution	Maximum Distance	Alpha	Beta	Mean	SD
Cycle	949583	Lognormal	30	1.5	1.09	6.28	8.64
Bus	168791	Weibull	100	0.97	0.97	29.53	41.86
Car	73165	Lognormal	200	1.79	1.39	13.76	26.11
2W	553501	Lognormal	100	1.36	1.19	7.45	12.84
Walk	1259203	Exponential	10	0.24	0.71	2.18	2.24
IPT	174579	Lognormal	100	1.76	1.01	9.36	14.3
Train	112869	Weibull	200	0.75	0.75	55.74	76.29

TABLE III: Data Description for Jharkhand: Using Author Data

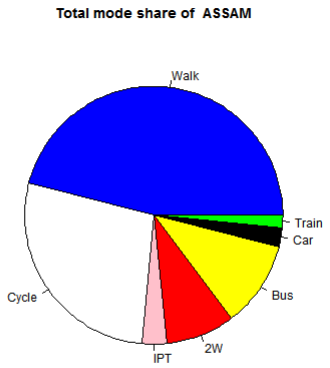

Fig. 1: Pie Chart : Assam

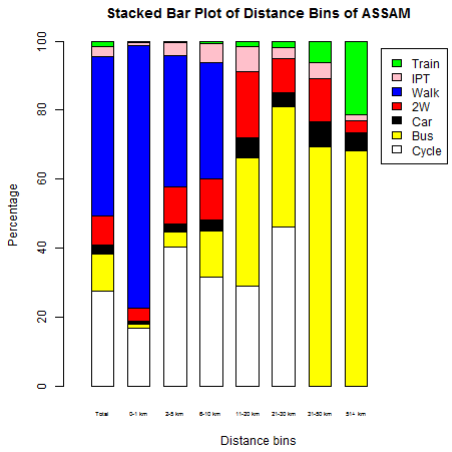

Fig. 4: Bar Chart : Assam

Data Name	Values/Results for Assam
Cluster Number	2
Most Popular Mode	Walking
Most Travelled Mode (from Cluster Analysis)	Bus
Fatality Rate	7.1

Fig. 2: Brief Summary Table for Assam

	Walk	Cycle	IPT	2W	Bus	Car	Train
Total Commuters	1736610	1041780	116593	327518	408509	90610	57863
Mean Distance	1.541037	4.670268	8.080387	7.241024	27.66515	14.59948	25.52317
SD of Distance	2.013192	7.170242	13.45303	13.07983	37.94777	29.05108	39.05044
Distribution	Exponential	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal	Weibull
Mode Commute Share	0.093079	0.152646	0.032788	0.082528	0.481127	0.046042	0.111789

Fig. 3: Modewise descriptive details for Assam

Assam is a very interesting state. Not only Culturally, but even from our perspective. Geographically, it has its share of hilly terrain, plateau structures as well as a river basin plain. The population is mainly concentrated near the plain along the river Brahmaputra. However, there is a significant contribution of rural parts to commute distance. (To access municipal and private services).

However, as is evident, the topography prevents people, specifically local villagers, to access modes beyond bus, cycle and walk. Hence, a majority is seen in those modes across all distances as well as in the aggregate. A more careful analysis would tell us that beyond a certain distance, Bus becomes not only a favoured mode but also the dominating one, akin to the role played by cycle for shorter distances. Presence of IPT and Train is negligible.

One visual anomaly or better framed, misdirection is that

bus seems to have a relative majority(sometimes absolute) in most sections of distance bins in the bar plot. However, this is nothing to be surprised of. The total can be thought of **Weighted** averages of these bars. Since the majority of the population has a commute lesser than 10km, the volume of bus travellers is overshadowed by cyclists and pedestrians.

§Haryana

Like Assam, we now look into another state of interest - Haryana. The Varying nature of distributions due to the vastly different geography and demography will become imminent here.

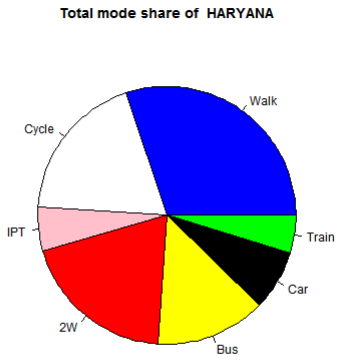

Fig. 5: Pie Chart : Haryana

Data Name	Values/Results for Haryana
Cluster Number	1
Most Popular Mode	Walking
Most Travelled Mode (from Cluster Analysis)	Bus
Fatality Rate	12.8

Fig. 6: Brief Summary Table for Haryana

	Walk	Cycle	IPT	2W	Bus	Car	Train
Total Commuters	957738	599093	175521	617205	436386	240413	151963
Mean Distance	2.010759	5.997629	10.16142	8.431051	4.461808	17.68717	25.09366
SD of Distance	2.322035	8.460643	15.66851	14.64966	7.186477	31.32548	35.54291
Distribution	Exponential	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal	Weibull
Mode Commute Share	0.046409	0.07828	0.042991	0.125434	0.40924	0.102528	0.195118

Fig. 7: Modewise descriptive details for Haryana

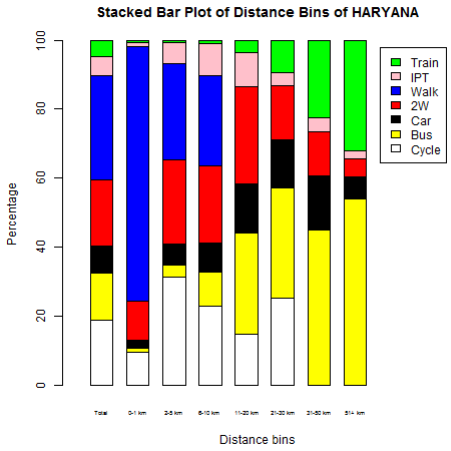

Fig. 8: Bar Chart : Haryana

Being Less populated than Assam yet slightly more densely populated, this is a very interesting example for comparison purpose.

Even though Haryana has a variety in geography, the majority of the population is centred around plains. Moreover, the population is characterised with a lot of daily workers. However, the variation of location and presence of hilly structures do not affect most of the population, as they only engage in land commute.

The locality of jobs and employment makes walking as a majority. However, Cycling and 2W are also pretty popular options for the same task. So the majority of walking is not as domineering as it was for Assam.

Also, the lack of physical undulation creates makes it possible for other modes like Train, IPT and Car to have a significant contribution even in lower distance bins. (Mainly for long-distance commuters, daily workers going to Chandigarh/NCT and economically better off population).

For interested readers, a complete collection of all graphs are present in Google Drive and the link has been shared later on. We shall also discuss two strong variations(outliers) observed in a later section.

C. Samples of Mode Distance Share and Cluster Output

Greener the shade, is lesser than mean
Redder the same, it is relatively more
Better is a state, if fatality is green
Deeper a shade, further away does it sore

We now look at the mode distance share and perform a k-means cluster analysis on both our obtained commute distance as well as Author's commute distance.

We shall study the clusters generated by our data, explain them and try to explain how geographically this is indeed not disjoint from reality.

Let us look at the clusters and the averages for our data.

State	Walk	Cycle	IPT	2W	Bus	Car	Train	Fatality Rates	Cluster
Cycle and Train (2W)									
ANDHRA PRADESH	0.06933	0.078803	0.080005	0.162655	0.408394	0.046421	0.154391	17.9	1
HARYANA	0.046409	0.07828	0.042991	0.125434	0.40924	0.102528	0.195118	12.8	1
JHARKHAND	0.102625	0.222942	0.06109	0.154161	0.186343	0.037638	0.235202	9.6	1
MADHYA PRADESH	0.107062	0.145343	0.029726	0.200241	0.275864	0.043099	0.198665	12.4	1
Bus and 2W									
ASSAM	0.093079	0.152646	0.032788	0.082528	0.481127	0.046042	0.111789	7.1	2
CHANDIGARH	0.044144	0.253226	0.040204	0.243256	0.200647	0.198179	0.020344	12.3	2
DADRA & NAGAR HAVELI	0.094484	0.086602	0.135041	0.245645	0.311567	0.070049	0.056611	17.3	2
GOA	0.03712	0.015056	0.022733	0.296929	0.521668	0.090957	0.015538	12.2	2
GUJARAT	0.0723	0.096382	0.117289	0.28202	0.258421	0.079155	0.094433	22.4	2
HIMACHAL PRADESH	0.111786	0.016052	0.015761	0.077164	0.70035	0.056469	0.022417	18.8	2
JAMMU & KASHMIR	0.079086	0.015233	0.028055	0.045213	0.678078	0.076607	0.077728	11.4	2
KERALA	0.056744	0.021998	0.020241	0.107819	0.471638	0.043844	0.277718	15.3	2
MANIPUR	0.082812	0.096189	0.112321	0.141687	0.44213	0.093102	0.031759	12.4	2
PUDUCHERRY	0.037762	0.120295	0.026099	0.332392	0.409545	0.050296	0.023611	18.8	2
PUNJAB	0.071931	0.198793	0.032624	0.204138	0.344878	0.066592	0.081043	14.2	2
RAJASTHAN	0.061154	0.057258	0.031653	0.154537	0.458339	0.048595	0.188463	13.6	2
UTTAR PRADESH	0.060785	0.171947	0.039084	0.13348	0.203356	0.040362	0.350986	7.5	2
UTTARAKHAND	0.094397	0.114228	0.05161	0.161788	0.398255	0.108889	0.070832	8.8	2
Walk-IPT-Car									
ARUNACHAL PRADESH	0.255927	0.04894	0.043967	0.145579	0.209191	0.269917	0.026479	9.7	3
DAMAN & DIU	0.135194	0.060962	0.053204	0.18086	0.467026	0.028678	0.074076	10	3
MEGHALAYA	0.119192	0.012938	0.115405	0.045276	0.460696	0.236009	0.010484	5.5	3
MIZORAM	0.15581	0.018544	0.073652	0.142704	0.382372	0.205883	0.021037	7	3
NAGALAND	0.178086	0.025758	0.160845	0.073407	0.258726	0.284926	0.018252	7.3	3
SIKKIM	0.189067	0.007021	0.170785	0.026401	0.094817	0.48286	0.029049	12.1	3
TRIPURA	0.138463	0.174726	0.09263	0.098169	0.337329	0.132762	0.025922	6.8	3
Cycle and Train									
BIHAR	0.102669	0.179572	0.033369	0.109567	0.133375	0.030742	0.410707	4.7	4
INDIA	0.064514	0.098815	0.04191	0.145879	0.336106	0.059543	0.253232	16	4
MAHARASHTRA	0.054276	0.04463	0.046969	0.145663	0.196852	0.047498	0.464112	11.6	4
WEST BENGAL	0.05715	0.135737	0.016941	0.043345	0.256256	0.027264	0.463307	6.3	4
Cycle-Bus-2W									
CHHATTISGARH	0.070194	0.244636	0.025608	0.290993	0.192569	0.053261	0.12274	12.4	5
KARNATAKA	0.07217	0.044394	0.058278	0.15295	0.48896	0.087649	0.0956	6.8	5
NCT OF DELHI	0.047402	0.071867	0.041785	0.206314	0.353938	0.203349	0.075344	2.3	5
ODISHA	0.082365	0.222252	0.025335	0.196758	0.279795	0.036172	0.157323	9.2	5
TAMIL NADU	0.040891	0.060118	0.021832	0.154479	0.542189	0.052926	0.127566	21.7	5
Mean	0.090776	0.09977	0.057113	0.15616	0.357354	0.104067	0.134761	11.59411765	

Fig. 9: Cluster Analysis for Our Commute Distance

Cluster	Walk	Cycle	IPT	2W	Bus	Car	Train	Fatality Rates
1	0.081356	0.131342	0.053453	0.160623	0.31996	0.057422	0.195844	13.18
2	0.071256	0.101136	0.050393	0.179186	0.42	0.076367	0.101662	13.72
3	0.167391	0.049841	0.101498	0.101771	0.315737	0.234434	0.029328	8.34
4	0.069652	0.114689	0.034797	0.111114	0.230647	0.041262	0.397839	9.65
5	0.062604	0.128653	0.034568	0.200299	0.37149	0.086671	0.115715	10.48

Fig. 10: Means of Our Clusters

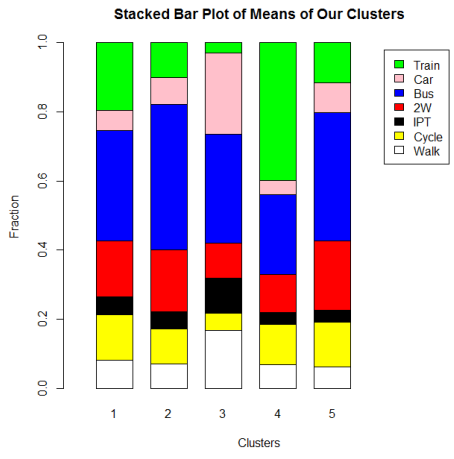

Fig. 11: Stacked Bar Plot : Cluster Means of Our Data

First, let us explain the colour coding.

As one can see, the means are listed at the bottom row of the table. We have coded red if the share of some state for that mode is more than the average share. The colour is gradually toned to white, the mean colour and then again increased to a deeper shade of green.

For the nomenclature of a cluster, we have used the following convention - We look at the mean of the cluster 10. The cells shaded orange are the ones above the mean by a significant factor, say at least 0.01. Marginal differences have been ignored to make better sense. However, had we faced difficulty with the categorization, we have included modes sufficiently close to the mean inside brackets to distinguish.

Note that similarity of colours of a mode with fatality for a state represents a positive correlation. This tells that that is indeed not so safe as a mode of travel. Negative correlations are better as that makes the mode feasible as well as a better choice.

We now try to explain how geography and demography plays a key role in clustering.

- Cluster 1 (Cycle and Train; 2W): These states are mainly composed of mainlands. Also, there are daily commuters who travel considerable distance daily to arrive at the place of work via public or pedestrian transportation. Plain lands have permitted in a feasible development of train, making it a cheaper and surer option compared to bus or IPT for transportation. Population density is also pretty high. That accounts for an increased fatality rate in some instances and the preference of walking and cycling for local commute.
- Cluster 2 (Bus and 2W): Majority of states are in this cluster. The funny thing is that they have a huge variation in demography and geography. However, a closer look tells that there are primarily three subcategories - Hilly (Assam, Uttarakhand, Chandigarh, Manipur, Uttar Pradesh and Kashmir), UTs and some larger states. Of course, we cannot consider them to be of similar geography, but they do have trends explainable. 2W Domination is found in smaller units whereas cycle domination is found in locations where it serves as an easier and economic mode due to lack of other commutes. Interestingly, bus domination is almost unanimous
- Cluster 3 (Walk-IPT-Car): Mainly hilly states are there in this category, primarily northeast. Here is an instance of significant similarity of geography. Train and Bus are not so prominent in most of these states. Even Cycle as a secondary mode has a considerable contribution.
- Cluster 4 (Cycle and Train): This is surprisingly similar to the first category, However, there is strong domination for those modes, unlike the first cluster, which makes it distinct. Also, fatality rates on average are lesser. However, the explanation is similar to the first cluster. The clear dominance is due to the fact train systems were pretty old in these states, hence that has developed pretty well.
- Cluster 5 (Cycle-Bus-2W): Finally, this category is also somewhat similar geographically to Clusters 1 and 4. However, fatality is pretty high and, especially in Tamil Nadu. These consist of mainly coastal states. Here buses are the primary public transport. Though the train is pretty popular as a mode in most states, Karnataka reduces this. Thus it is a secondary mode.

As stated earlier, local minima (even global minima) selection has some randomness involved. So a rerun might give slightly different clusters. Also, many elements from

different clusters seem to have similar properties both data-wise and physically, whereas some units in a cluster differ considerably. This can be accounted as an error associated with the randomness of `kmeans` function to reach the minima.

To Conclude, let us look at a bar graph of the means of clusters for a better understanding. Also, following that, we have posted the cluster analysis of distances found from Author's Mean. One can make a similar analysis.

Cluster	Walk	Cycle	IPT	2W	Bus	Car	Train	Fatality Rates
1	0.079193	0.104169	0.054374	0.182448	0.410637	0.079382	0.089797	12.22
2	0.06872	0.132972	0.034091	0.108014	0.19746	0.036466	0.422278	7.53
3	0.072136	0.085626	0.05542	0.189451	0.455839	0.120836	0.020692	13.33
4	0.183471	0.054998	0.108376	0.097252	0.256487	0.27527	0.024148	8.58
5	0.073775	0.115712	0.041619	0.155936	0.353215	0.05223	0.207514	13.35

Fig. 12: Means of Author Clusters

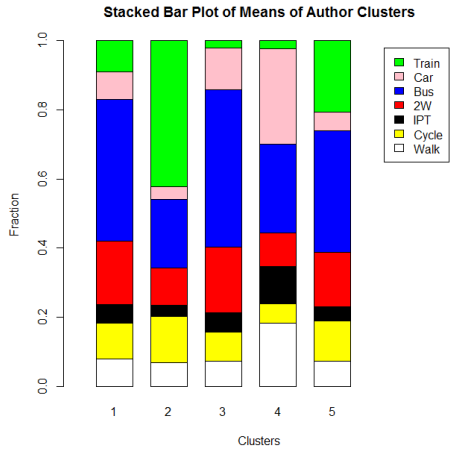

Fig. 13: Stacked Bar Plot : Cluster Means of Author Data

State	Walk	Cycle	IPT	2W	Bus	Car	Train	Fatality Rate	Cluster
2W and Bus									
ASSAM	0.093079	0.152646	0.032788	0.082528	0.481127	0.046042	0.111789	7.1	1
CHHATTISGARH	0.070194	0.244636	0.025608	0.290993	0.192569	0.053261	0.12274	12.4	1
DADRA & NAGAR HAVELI	0.094484	0.086602	0.135041	0.245645	0.311567	0.070049	0.056611	17.3	1
DAMAN & DIU	0.135194	0.060962	0.053204	0.18086	0.467026	0.028678	0.074076	10	1
GUJARAT	0.0723	0.096382	0.117289	0.28202	0.258421	0.079155	0.094433	22.4	1
JAMMU & KASHMIR	0.079086	0.015233	0.028055	0.045213	0.678078	0.076607	0.077728	11.4	1
KARNATAKA	0.07217	0.044394	0.058278	0.15295	0.48896	0.087649	0.0956	6.8	1
NCT OF DELHI	0.047402	0.071867	0.041785	0.206314	0.353938	0.203349	0.075344	2.3	1
PUNJAB	0.071931	0.198793	0.032624	0.204138	0.344878	0.066592	0.081043	14.2	1
TAMIL NADU	0.040891	0.060118	0.021832	0.154479	0.542189	0.052926	0.127566	21.7	1
UTTARAKHAND	0.094397	0.114228	0.05161	0.161788	0.398255	0.108889	0.070832	8.8	1
Cycle and Train									
BIHAR	0.102669	0.179572	0.033369	0.109567	0.133375	0.030742	0.410707	4.7	2
MAHARASHTRA	0.054276	0.04463	0.046969	0.145663	0.196852	0.047498	0.464112	11.6	2
UTTAR PRADESH	0.060785	0.171947	0.039084	0.13348	0.203356	0.040362	0.350986	7.5	2
WEST BENGAL	0.05715	0.135737	0.016941	0.043345	0.256256	0.027264	0.463307	6.3	2
2W-Bus-Car									
CHANDIGARH	0.044144	0.253226	0.040204	0.243256	0.200647	0.198179	0.020344	12.3	3
GOA	0.03712	0.015056	0.022733	0.296929	0.521668	0.090957	0.015538	12.2	3
HIMACHAL PRADESH	0.111786	0.016052	0.015761	0.077164	0.70035	0.056469	0.022417	18.8	3
MANIPUR	0.082812	0.096189	0.112321	0.141687	0.44213	0.093102	0.031759	12.4	3
MEGHALAYA	0.119192	0.012938	0.115405	0.045276	0.460696	0.236009	0.010484	5.5	3
PUDUCHERRY	0.037762	0.120295	0.026099	0.332392	0.409545	0.050296	0.023611	18.8	3
Walk-IPT-Car									
ARUNACHAL PRADESH	0.255927	0.04894	0.043967	0.145579	0.209191	0.269917	0.026479	9.7	4
MIZORAM	0.15581	0.018544	0.073652	0.142704	0.382372	0.205883	0.021037	7	4
NAGALAND	0.178086	0.025758	0.160845	0.073407	0.258726	0.284926	0.018252	7.3	4
SIKKIM	0.189067	0.007021	0.170785	0.026401	0.094817	0.48286	0.029049	12.1	4
TRIPURA	0.138463	0.174726	0.09263	0.098169	0.337329	0.132762	0.025922	6.8	4
Cycle and Train (2W and Bus)									
ANDHRA PRADESH	0.06933	0.078803	0.080005	0.162655	0.408394	0.046421	0.154391	17.9	5
HARYANA	0.046409	0.07828	0.042991	0.125434	0.40924	0.102528	0.195118	12.8	5
INDIA	0.064514	0.098815	0.04191	0.145879	0.336106	0.059543	0.253232	16	5
JHARKHAND	0.102625	0.222942	0.06109	0.154161	0.186343	0.037638	0.235202	9.6	5
KERALA	0.056744	0.021998	0.020241	0.107819	0.471638	0.043844	0.277718	15.3	5
MADHYA PRADESH	0.107062	0.145343	0.029726	0.200241	0.275864	0.043099	0.198665	12.4	5
ODISHA	0.082365	0.222252	0.025335	0.196758	0.279795	0.036172	0.157323	9.2	5
RAJASTHAN	0.061154	0.057258	0.031653	0.154537	0.458339	0.048595	0.188463	13.6	5
Mean	**0.090776**	**0.09977**	**0.057113**	**0.15616**	**0.357354**	**0.104067**	**0.134761**	**11.59411765**	

Fig. 14: Cluster Analysis for Author's Commute Distance

D. Regression and Sensitivity Analysis

The results of the regression analysis are tabulated.

	Mean	SD	0.025q	0.5q	0.975q	Mode
(Intercept)	-10.230	0.621	-11.456	-10.231	-9.001	-10.232
ln(bus)	0.090	0.117	-0.141	0.090	0.320	0.091
ln(auto)	-0.098	0.127	-0.350	-0.098	0.154	-0.098
ln(car)	0.063	0.118	-0.170	0.063	0.295	0.063
ln(walk)	**-0.250**	0.148	-0.542	-0.250	0.043	-0.250
ln(bicycle)	-0.211	0.087	-0.383	-0.211	0.038	-0.211
ln(X2w)	**0.447**	0.130	0.190	0.447	0.704	0.446

Results of Regression Analysis

Interpretations and Explanations:

- The sign of the mean corresponding to each mode of transport is an indicator of the effect of the mode on the annual fatality count.
- One can observe that a positive (negative) value of the mean implies that there is a positive (negative) correlation between the commute distance corresponding to a mode and the expected annual fatality count; Also, a higher magnitude of the mean implies a higher magnitude of the correlation.
- In the table, red (green) color is used the respective mean of each mode with positive (negative) correlation with the fatality risk.
- Bus has a mixed effect since the coefficient is positive, but it also contributes to walking mode which has a negative correlation with the fatality risk
- IPT has a negative correlation. This can be reasoned from the fact that they have an enclosure protecting the passengers and have a smaller engine size compared to cars and buses, thereby posing less danger to other road users.
- So, all Public Transport modes, directly or indirectly contribute to reduced fatality risk.
- Two wheeler mode has the highest correlation with the expected annaul fatality count possibly because the mode is unsafe to the rider as well as cyclists and pedestrians.
- Car, bicycle and walk modes have positive, negative and negative correlations with the fatality risk respectively. These can be justified by similar arguments.

The results of the sensitivity analysis are tabulated.

	Mean	SD	0.025q	0.5q	0.975q	Mode
(Intercept)	-10.230	0.621	-11.456	-10.231	-9.001	-10.232
ln(bus)	0.029	0.109	-0.187	0.029	0.244	0.029
ln(auto)	-0.122	0.132	-0.383	-0.123	0.138	-0.123
ln(car)	0.034	0.119	-0.201	0.034	0.269	0.034
ln(walk)	-0.115	0.098	-0.309	-0.115	0.078	-0.115
ln(bicycle)	**-0.209**	0.098	-0.403	-0.209	-0.015	-0.209
ln(X2w)	**0.431**	0.145	0.145	0.431	0.717	0.431

Results of Sensitivity Analysis - Case 1 (No walking distance)

	Mean	SD	0.025q	0.5q	0.975q	Mode
(Intercept)	-10.151	0.597	-11.331	-10.152	-8.970	-10.153
ln(bus)	0.111	0.121	-0.130	0.111	0.349	0.112
ln(auto)	-0.093	0.127	-0.345	-0.093	0.159	-0.093
ln(car)	0.069	0.118	-0.163	0.069	0.301	0.069
ln(walk)	**-0.285**	0.163	-0.607	-0.286	0.037	-0.286
ln(bicycle)	-0.217	0.085	-0.385	-0.217	0.050	-0.217
ln(X2w)	**0.457**	0.127	0.205	0.457	0.709	0.456

Results of Sensitivity Analysis - Case 2 (1.5 km of walking distance)

Interpretations and Explanations:

- In accordance with expectations, the correlations corresponding to Public Transport modes with the expected fatality count decrease in the first case (No walking distance) and increase in the second case (1.5 km of Walking distance)
- The correlation corresponding to Walk mode increases (decreases in magnitude) in the first case and decreases (increases in magnitude) in the second.

E. Mode Shift

The following plots were obtained on completion of the procedures mentioned in the previous section.

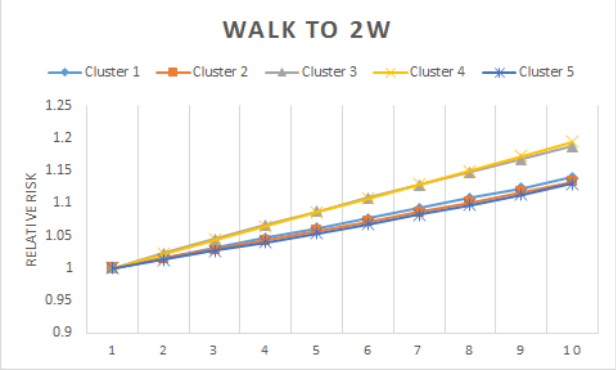

Fig. 15: Walk to Two Wheeler Shift - Line Plot

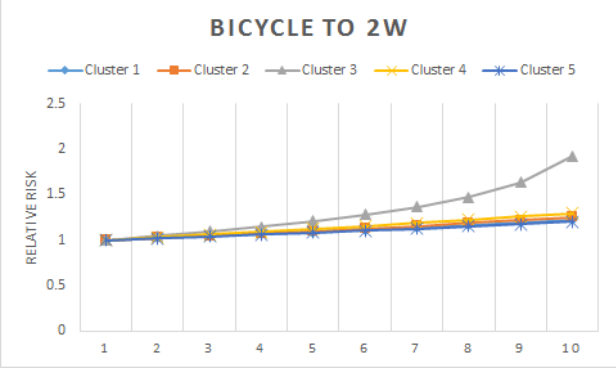

Fig. 16: Bicycle to Two Wheeler Shift - Line Plot

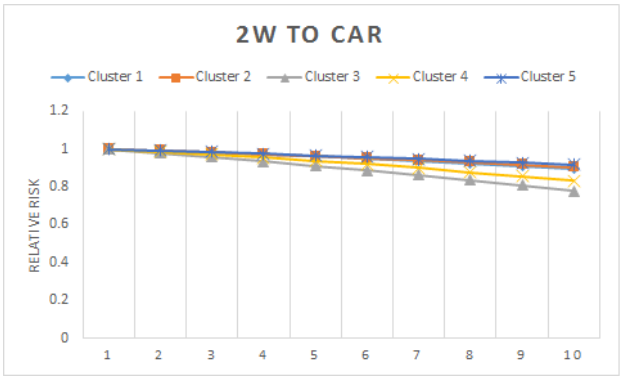

Fig. 17: Two Wheeler to Car Shift - Line Plot

Interpretations and Explanations:

- **Walk to Two Wheeler:** The Relative Risk increases for all clusters. The increase is larger for Cluste 3 and Cluster 4. This is due to the very low mode shares of Two wheeler mode in these two clusters.
- **Bicycle to Two Wheeler:** The Relative Risk increases. The increase is substantially large for Cluster 3 due to low mode share of Bicycles.
- **Two Wheeler to Car:** The shift leads to a decrease in Relative Risk as according to the model, Two Wheeler mode is associated with the highest risk.

VI. Discrepancies in Source Data and Possible Errors in Ours

We tried to be as discreet and descriptive as we can. So we reran our codes and functions multiple times. However, we faced some difficulty and also some considerable deviations were spotted. They are stated in brief over here

A. Some Candid Discrepancies in source data/document

On a cursory glance at the supplementary data provided by the author for modelling the distance decay functions, it is easy to see that for the Weibull distribution, something seems off. Specifically, all the parameter-pairs have only one value. We believe that this is an error and needs to be fixed.

B. Formal Assumptions made by the author with vague accountability and reasoning (at our level)

Based on un-verbalised knowledge in the author's domain of expertise, the right distributions for modelling Distance Decay functions are Exponential, Log-Normal and Weibull. We lack the experience and knowledge to justify these choices. Thus, we have used the same distribution as used by the author to model the Distance Decay Functions.

The author mentions that based on a preliminary analysis, the parameter values for the fitted distributions lie in small ranges. We are in the dark about any possible method that would give us that conclusion. So, we have used the non-erroneous values given by the GRG solver.

C. Some discrepancies between the Author's calculations and ours

We made an attempt to replicate all the regression models proposed. However, in 'Model 4' (this wasn't discussed in the presentation or this report), the author uses 'Population Density' as a component of the vector of explanatory variables. But, when the coefficient corresponding to this component is estimated using R-INLA, we get '0' as the mean. We opine that this happens because every component of X_n in the observed data set is less than '20 units' whereas, the average population density is of the order of 10^4 sq. km [9]. The author, however has obtained '0.039' as the mean value of the same. Given that we had obtained a decent match with the author's calculations for the other three models, we believe that this is an issue with the definition[10] or the results given.

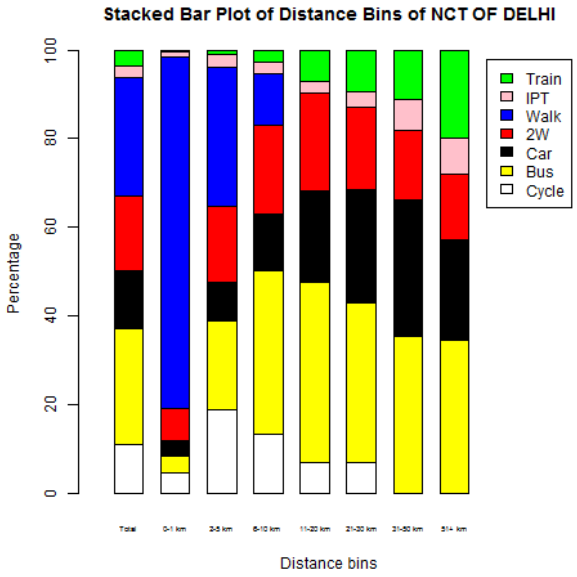

Fig. 18: Means of Our Clusters

D. Assumptions made to extend on the study

There were some natural assumptions made to extend the studies. Even though there were some unexplainable outliers in our data, we thought of it as sampling error to account for it, as did the author. An example would be the following :

Quoting the author "In case of cycle,it was observed that in most states the counts in 21–30 km bin are more than,or in some cases almost the same as, the preceding bin of 11–20km. Note that both bins are of equal size (10km). There are some exceptions,such as Delhi,where number of cycle trips in the last bin(21–30 km) are 30% of those in the preceding bin(11–20). Any distribution for cycle distance is likely to have a negative slope at this distance range (> 10 km) Therefore, with the two bins of equal size(10km), it is not possible to have more number of trips in 21–30km range than 11–20 km. There is some clear discrepancy in the data"

Not just this, there were some other instances akin to the above as well. The Author has suggested merging of these bins, but not only shall it lead to loss of information, but the structure won't be compatible with our analysis. Hence, we have not performed such rectification. What follows are the bar plot of Delhi to back up the Author's stance.

We have also assumed that for extending our study to rural, urban structure, these are the only three possible distributions and followed the Author's Method word to word. Some experience would have been helpful over here.

E. Variation on retrying codes and R-based errors

In this brief section, we discuss how re-running codes in R can create different outputs each time.

The Means, SD and other computation like mode share, mode distance share, commute distance did not produce variable results as such. However, a significant variation was found while rerunning R-codes for the K-means analysis. A possible cause might be the existence of multiple minima with negligible difference in WCSS. There might also be an error due to difference of algorithms used by us and the author as it is not mentioned explicitly (We used Hartigan-Wong, the default in R).

Similarly, another optimisation problem which might produce slight variation upon retry is regression. However, in this case, the variation was seen to be minimal.

These might be causes of errors. As an example, we hereby post three graphs obtained from the same run to determine knees in number of cluster.

F. Parts where we encountered problems and took measures (possibly leading to minor errors)

The author does not mention the version of Microsoft Office used for Optimisation. Our analysis was done on the Office365 ProPlus Version 2002 Build 12527.21236. Thus minor deviations in the parameter values are to be expected due to presumed improvements in the GRG solver algorithm.

Also, the author states that the initial values for the optimisation are to be 1 for both alpha and beta. On doing that, all instances of the Weibull distribution gave a #VALUE error.

[9] `othervariables.xlsx` contains all the additional data pertaining to Models 2, 3 and 4

[10] The author never explicitly defined in the paper, the vector of explanatory variables used for the models. So, we had to infer the particulars from 'Table 3'. Hence, although it was mentioned as just 'population density' , we do not know if the variable (Population density) was used in some other form (for instance logarithmic, scaled ...) which might yield the desired result.

On tinkering with the parameter values, we observed that the beta values much larger than 1 do not give that error. Thus, in all these cases, we started the GRG algorithm with alpha=1 and beta=10. We have also bounded beta above by 20 for the algorithm to run without hindrance. This might not be a feasible assumption, but without it, the optimiser seemed to be returning high values of beta, causing problems in the calculation of mean and sd.

A closer look and inclusion of more distributions might have proven to be useful and more informative. But this is beyond our scope hence we have decided to overlook it.

G. Possible Normalisation and other variation

In his Data, the author has included something known as Normalisation Constant, the application of which was neither discussed in the paper nor is it obvious. We could not figure out what it represented. Initially, we thought it might represent some division or multiplication factor. However, though everything was lesser than 1, some values were pretty small, even 0. So we dropped our idea and ignored the same. This might have been useful somehow.

Also, there has been some instance of significant variation. We can think of it as due to some sampling error, additional unaccounted constraints implemented by the author or some oversight in the documentation. However, with whatever we had, we tried to implement it in R and have successfully verified that on the face of it, our codes are running correctly.

There were also different modes and options for different functions used. We tried to make the best choice or followed the default option. This might have caused some deviation from the Author's Data.

VII. Strengths, a generalisation of the model and possible reusability

- This paper reports the use of census data to develop an injury prediction model accounting for exposure of all road users, unlike previous.
- This is the first such study in India and the methods can be applied to model injuries at the city or district level.
- We also get a rough estimate of how mode shifts and increasing privatisation/mechanization can have concerning road deaths on a pro-rata basis
- Inter alia, it also frames an ecological model with modifiable areal units for a stronger analysis given sufficient data. Thus, our model is more robust and reusable vis-a-vis the existing ones
- The ecological character of the model can be applied to various situations. Consider, for example, a profitable hierarchical organisation with various categories of workers. For each category, we divide the weekly working hours into bins and record the corresponding counts of workers. Further, for each category of workers, there is an upper limit on the weekly working hours. Under these assumptions, (and, probably, independence of working hours across different categories), it is possible to model the net profit earned by the company in terms of the total working hours of various categories of workers.

  This model, of course, has some downsides. There are more significant factors like the payments and work output of various categories of workers. Further, in some cases, one might not even have working hours as an explanatory variable- for instance, when you are hired by the company to achieve targets with time restrictions. So, for such posts, variation in working hours doesn't change the profits.

  , However, if we could fit in all conditions assumed in our model, the analysis would be greatly simplified by applying our techniques.

VIII. Some Extra Work!!

Throughout the project, our group has tried to create and add extra outputs, like using own R-codes, creating graphs beyond the ones presented, re-interpreting data and output. Some (but not exhaustive) Past Instances Are -

1) Pie and Stacked Bar Charts of Different State Mode Shares
2) Stacked Bars for Cluster Analysis
3) Using our R functions for distribution and Mean
4) Using VBA to automate the calculation
5) A Relation of Geography and Demography with Mode Share as well as Cluster Analysis.

In this section, we again perform some mini activities to understand the data better.

A. Checking Independence of Mode and Distance Bin

This is a fun activity which uses concepts from our present course and tries to analyse that for the three modes having maximum distance 100 km, namely Bus, 2W and IPT, whether or not the distribution of the population in distance bins is independent of mode.

As our data consists of huge units, direct analysis of independence would be futile. So we instead take people per bin in 1000 to the analysis.

We use the code present in `chisqind.r`. Let us see how the analysis is done.

Raw Data is drawn from any source table with untampered data. After this, the `checkind` function goes state by state forming sub tables with columns as distance bins and rows as modes. We round off after dividing each entry by a thousand.

Next, a Two-Way Independence Test(also called a Chi-Squared Independence Test) is performed on this subtable.

Based on Row Sums and Column Sums, we calculate the expected value of a cell if it were independent. After that, we perform a Chi-Squared Test for independence using the natural test statistic. We assume that all assumptions needed hold and also compare our test with different rejection values, basis which we perform rejection and hypothesis testing. The process is explained in the following brief example

Consider the following table :

	A	B	C	Row Sums
X	v_{11}	v_{12}	v_{13}	r_1
Y	v_{22}	v_{22}	v_{23}	r_2
Col Sums	c_1	c_2	c_3	Total=N

Then Expected value for the (i,j)-th cell assuming independence of rows and columns is $e_{ij} = \frac{r_1 c_j}{N}$.

The test statistics is defined as $\chi^2 = \sum_{i,j} \frac{(v_{ij}-e_{ij})^2}{e_{ij}}$

With proper assumptions, it turns out χ^2 has the distribution of $Chi-square$ with degrees of freedom = $(r-1)\times(c-1) = 1\times 2 = 2$, where r =number of rows and c =number of cols.

We do the rest of the analysis on these test statistics. After the code is run, the data is exported and missing links like labels and proper representation are added. Whatever we achieved has been presented in the image 19.

Sl No.	State	Level of Type I error (Alpha)				
		0.1	0.05	0.02	0.01	0.001
1	ANDHRA PRADESH	×	×	×	×	×
2	ARUNACHAL PRADESH	✓	✓	✓	✓	✓
3	ASSAM	×	×	×	×	×
4	BIHAR	×	×	×	×	×
5	CHANDIGARH	✓	✓	✓	✓	✓
6	CHHATTISGARH	×	×	×	×	×
7	DADRA & NAGAR HAVELI	✓	✓	✓	✓	✓
8	DAMAN & DIU	✓	✓	✓	✓	✓
9	GOA	✓	✓	✓	✓	✓
10	GUJARAT	×	×	×	×	×
11	HARYANA	×	×	×	×	×
12	HIMACHAL PRADESH	×	×	×	✓	✓
13	INDIA	×	×	×	×	×
14	JAMMU & KASHMIR	×	×	×	×	✓
15	JHARKHAND	×	×	×	×	×
16	KARNATAKA	×	×	×	×	×
17	KERALA	×	×	×	×	×
18	MADHYA PRADESH	×	×	×	×	×
19	MAHARASHTRA	×	×	×	×	×
20	MANIPUR	×	×	×	×	×
21	MEGHALAYA	×	×	✓	✓	✓
22	MIZORAM	✓	✓	✓	✓	✓
23	NAGALAND	✓	✓	✓	✓	✓
24	NCT OF DELHI	×	×	×	×	×
25	ODISHA	×	×	×	×	×
26	PUDUCHERRY	×	×	✓	✓	✓
27	PUNJAB	×	×	×	×	×
28	RAJASTHAN	×	×	×	×	×
29	SIKKIM	✓	✓	✓	✓	✓
30	TAMIL NADU	×	×	×	×	×
31	TRIPURA	×	×	×	×	×
32	UTTAR PRADESH	×	×	×	×	×
33	UTTARAKHAND	×	×	×	×	×
34	WEST BENGAL	×	×	×	×	×

Fig. 19: Test of Independence

Here a tick represents the possibility of independence whereas a cross represents a rejection of this possibility.

Now let us briefly analyse our data. In most instances, it has been observed that the data is either not independent for each alpha for given confidence or it has a possibility of being independent for all of them. But is this possibility real?

All of these units have one property in common - All of them are small states, both size and population-wise. If we closely analyse the chi-squared, it generally turns out to be very less, the cause being that it is not wise enough to consider per thousand population for these units. This gives us the independence, so at smaller scales, like per 100 estimates, would have given us better results here.

However, the interesting part appears in the states of Himachal Pradesh, Jammu and Kashmir, Meghalaya and Puducherry.

For Puducherry, even though the dataset is small, the variation is tremendous, thereby rejecting independence for lower confidence despite fewer numbers.

For the rest, as well as Mizoram, Nagaland and Arunachal Pradesh this hints at a possibility of independence. Chi-Square is not small as the data is not per se negligible as the data sets are relatively larger. Yet, with appropriate confidence, we have achieved the possibility of independence.

It is interesting to note that the units giving positive result for given alphas are either very small or hilly (Like JnK, HP or North-East). This leads us to conjecture the following - Given Lesser Transport Facility and somewhat lower to moderate numbers, will at sufficient yet not negligible confidence, a scope of independence appear?

B. Extending Mode Share data and distribution to Rural/Urban Framework

Our primary analysis (before regression) can be extended to just study the Rural or Urban set up.

In this activity, we look at individually the urban and rural set up for our home states West Bengal, Tamil Nadu and Uttar Pradesh.

As done earlier, we start with the assumption that Walking is distributed Exponentially and Cycling is Distributed Log-normally.

We assume the maximum distances to be the same as the total data. Now we start by assuming lognormal distribution in the other modes. After Computing alpha and beta, we calculate the Pearson Correlation for these two data. If it is not sufficiently large, we change the distribution to Weibull and optimize accordingly.

After this, we process to calculate the mean and variance, as we did for the data for the total state population.

The results are presented below. Had we had access to the data for road fatality of solely either of these, we could have constructed a separate model or might have tested how accurate our initial model is.

Pop. Type \ State		Tamil Nadu	Uttar Pradesh	West Bengal
Rural	Mean	5.723444	5.708113	5.067926
	SD	15.02764	14.90881	13.12503
Urban	Mean	5.661695	4.570913	7.207321
	SD	14.71216	11.90991	21.45313

Fig. 20: Mean and Variance for Bus in different modes

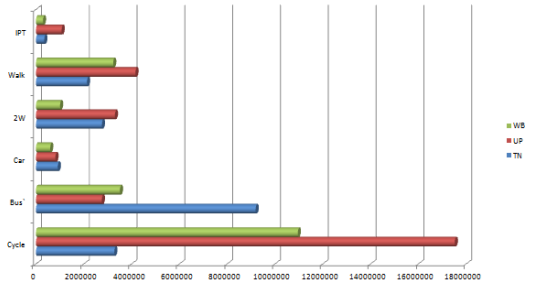

Fig. 21: Commute Distances - Rural

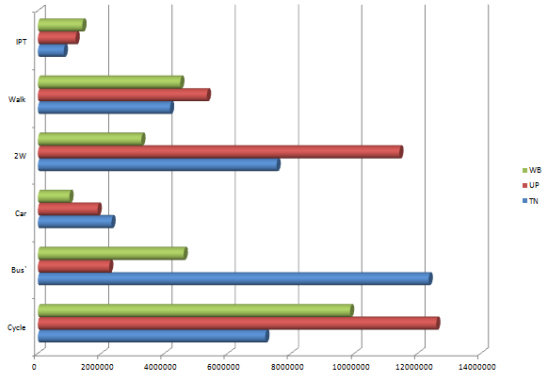

Fig. 22: Commute Distances - Urban

APPENDIX A
ALL R-CODES AND VBA IN BRIEF

Code used is present in the files attached.

DATASETS, REFERENCES AND BIBLIOGRAPHY

Sources of data

[1] Census Of India, 2011 Accessed online at the website https://censusindia.gov.in

[2] NCRB, 2011. Accidental Deaths & Suicides in India 2010. Accessed online. National Crime Records Bureau, New Delhi. http://ncrb.nic.in/.

[3] NCRB, 2012. Accidental Deaths & Suicides in India 2011. Accessed online. National Crime Records Bureau, New Delhi. http://ncrb.nic.in/.

[4] NCRB, 2013. Accidental Deaths & Suicides in India 2012. Accessed online. National Crime Records Bureau, New Delhi. http://ncrb.nic.in/.

[5] MoRTH, 2011. Review of the Performance of State Road Transport Undertakings (SRTUs)—Passenger Services for April 2010- March 2011. Transport Research Wing, Ministry of Road Transport and Highways, Government of India, New Delhi, India. MoRTH SRTUs 2010-11

[6] MoRTH, 2013. Basic Road Statistics of India 2011-12. Transport Research Wing, Ministry MoRTH 2011-12

[7] Open Government Data Platform India: https://data.gov.in/

[8] NRSC Web Portal (Bhuvan) : https://bhuvan-app1.nrsc.gov.in/thematic/thematic/index.php

[9] Elvik. R 2016 : Safety-in-numbers: Estimates based on a sample of pedestrian crossings in Norway

References

[10] Rue, H., Martino, S., Lindgren, F., 2009. INLA: Functions Which Allow to Perform a Full Bayesian Analysis of Structured (Geo-) Additive Models Using Integrated Nested Laplace Approximation. R Package Version 0.0 ed.INLA: Functions Which Allow to Perform a Full Bayesian Analysis of Structured (Geo-) Additive Models Using Integrated Nested Laplace Approximation. R Package Version 20.3.17

[11] Books/GitBooks consulted to use R-INLA: Bayesian Inference with INLA, Geospatial Health data: Modeling and Visualization with R-INLA and Shiny, Bayesian Regression Modeling with INLA

[12] Distance Decay Functions of Travel to Work http://refhub.elsevier.com/S0001-4575(17)30457-8/sbref0060

[13] Cluster Analysis in R was done by consulting https://techvidvan.com and https://www.rdocumentation.org

[14] F. Crénin, Truncated Weibull Distribution Functions and Moments (2015). https://ssrn.com/ abstract¼2690255.

[15] GRG Algorithm :https://www.refhub.elsevier.com/S2352-3409(18)31174-0/sbref6

[16] GRG-Nonlinear was used after consulting https://www.solver.com , https://www.dummies.com , and https://docs.microsoft.com/

[17] Visual Basic for Applications was used after consulting https://www.dummies.com and https://www.educba.com/vba-tutorial-for-beginners/