

The Enterprise Surveys of the Informal Sector in Cities in India, 2022

I. Introduction

This document provides information on the Enterprise Surveys of the Informal Sector (ESIS) data collected by the World Bank Group's (WBG) Enterprise Analysis Unit (DECEA) in several cities in India. The survey covers the following nine cities: Hyderabad, Telangana; Jaipur, Rajasthan; Kochi, Kerala; Ludhiana, Punjab; Mumbai, Maharashtra; Sehore, Madhya Pradesh; Surat, Gujarat; Tezpur, Assam; and Varanasi, Uttar Pradesh. The fieldwork was implemented by Nielsen (India) Pvt Ltd, and the data was collected between September 2021 and April 2022.¹ The ESIS in cities in India was made possible thanks to the cooperation with Omidyar Network India.

The primary objectives of the ESIS are to: i) understand demographics of the informal sector in the covered cities, ii) describe the environment within which these businesses operate, and iii) enable data analysis based on the samples that are representative at each city level.

This report briefly describes the sampling design of the data, the structure of the dataset as well as additional information that may be useful for data users.

II. Definition of Informality and Universe

The ESIS covers all informal businesses within a well-defined geographic area, typically a city. All eligible businesses are considered as forming the universe of inference for the survey. Informality is defined as all businesses that are not legally registered with the government and, therefore, are excluded from taxation. The exact details of these criteria vary by country. For the 2022 ESIS in cities in India, a business is considered informal if it does not have a business PAN or a GST number. Thus, the universe of the survey includes all businesses that meet this definition of informality, including all sectors of activity of any size. The universe excludes, however, any illicit or illegal activity.

III. Sampling Approach

This section briefly outlines the ESIS sampling approach (please see Aga et al. (2022) for further details). A challenge to conducting a representative survey of informal sector businesses is the lack of a proper sampling frame since these businesses are not registered and, therefore, they are almost always absent from official registries or any other potential sampling frame. The ESIS follows an area-based sampling methodology. Each city is overlayed with a spatial grid, dividing the area into squares of equal size.² In the case of the 2022 ESIS in cities in India these squares measured 150 meters by 150 meters. This spatial grid is prepared using a GIS software. Often, the boundaries of the city match a set of administrative boundaries. Nevertheless, when appropriate, a wider urban area (possibly outside of administrative boundaries) is used to best capture the extent of informal business activity in an area. The resulting map is reviewed and approved by the

¹ COVID-19 presented unique challenges during data collection. Safety of the field team and respondents was the highest priority of the WB team and Nielsen (India) Pvt Ltd. Fieldwork was thus relatively slow, with frequent pauses.

² This methodology can be applied to any geographic area. The ESIS is typically conducted within a city/well-defined urban area.

relevant WBG Task Team Leader (TTL) in consultation with the implementing contractor. For the purposes of the 2022 ESIS in cities in India, the shapefiles outlining administrative boundaries were obtained by Nielsen (India) Pvt Ltd from MapsofIndia.com. These outlines are given in Appendix A (dashed lines on the maps), along with the spatial grid used for each city.

Each square in the spatial grid, referred to as a block, constitutes a Primary Sampling Unit (PSU). The ESIS uses the Adaptive Cluster Sampling (ACS) method (Thompson, 1990, 1991), whereby an initial sample of starting blocks is randomly selected from the city's grid. See Appendix B for the number of starting blocks selected for each city. All encountered informal businesses within each starting block are then fully enumerated. ACS takes advantage of the expectation that informal businesses are geographically clustered. To do so, a threshold number of informal businesses is defined. All blocks that meet the threshold trigger full business enumeration of all surrounding blocks. This process of sample expansion continues until no blocks meet the set expansion threshold.³

Each sampled block (starting or selected through subsequent expansions) is thoroughly enumerated whereby basic information (such as sector, size, etc.) for all informal businesses located in the block is collected (listed businesses). Enumerators are instructed to list and account for businesses that refuse the listing exercise as well as those that are unavailable at the time of fieldwork (e.g., they are encountered during off-hours). In those cases, enumerators record the refusal/unavailability and record a few pieces of information by observation. A randomly selected subset of the available enumerated businesses is interviewed through a 25-minute questionnaire (“interview”). As such, the ESIS follows a two-stage sampling process, whereby a probability of selection is clearly defined by integrating selection probabilities from both stages for each interviewed business resulting in a sample that is representative of the informal sector at the city level.

For increased precision, the ESIS uses the stratified version of the ACS. After the spatial grid is created, blocks in the grid of each city are categorized into five strata: residential, commercial/industrial, mixed (commercial and residential), market center, and open areas.⁴ The stratification was based on local knowledge of cities by the respective field teams of Nielsen (India) Pvt Ltd. Note that blocks stratified as market centers are not selected using ACS but rather through simple random sampling (SRS).

IV. Survey Implementation

This survey was fully implemented using the World Banks’ Survey Solutions CAPI software. The selection for interviews was conducted in the field and in real time, via a random algorithm programmed in the CAPI so that enumerators do not have control over who gets selected. A detailed monitoring protocol was put in place during the data collection to ensure the integrity of the fieldwork and methodology. In addition to supervision through assigned supervisors, every enumerator records his/her path using a tracking software (Oruxmaps) installed

³ Different patterns of expansion can be implemented through this methodology. The ESIS typically activates all surrounding blocks of those meeting the expansion threshold. If the expansion process is considerably longer than expected, a TTL can force a stop to expansions.

⁴ There is a sixth category for blocks that are physically inaccessible. This category is excluded from the sampling.

on all CAPI tablets. Enumerators submit captured paths to a centralized server at the end of the enumeration of every block. This tracking path is checked to ensure that enumerators have fully covered the block assigned to them.⁵ This quality check was done at high frequency. In cases where the tracking path indicated a lower than acceptable level of effort in listing informal business, the enumerator was asked to re-survey the block. For detailed information about the total number of blocks covered, informal businesses enumerated, and interviews conducted in each city, please see Appendix B.

V. Database Structure

The main datafile is based on interviews conducted through a standardized questionnaire, which was piloted and reviewed before fieldwork, building on previous modules used by the World Bank's Enterprise Analysis Unit to survey informal businesses. For more information about the questionnaire, see Aberra et al. (2022). Together with the standardized questions, the questionnaire for the 2022 ESIS in cities in India includes a set of India-specific questions, all of which are marked with prefix "msc". For more details, please see the questionnaire provided as part of the survey documentation. The language in which the interview was conducted is given in variable *a1a*.⁶

The data contains a unique business identifier variable named *idstd*. Most variables in the dataset are numeric, except those with suffix "x", which are string variables. The variables *wweak*, *wmedian*, *wstrict* are the (consolidated from the two stages) sampling weights corresponding to the different weighting criteria (see next section for details). The sampling weights enable inferences to the population of informal businesses in each city and are strongly recommended to be used in data analysis. The variable *strata* indicates the sampling design stratum within each city. The variable *id_square* identifies businesses that have been interviewed in the same block.

All monetary questions record the answers in the local currency unit, the Indian Rupee.

VI. Sampling Weights⁷

To estimate population parameters, weights are applied to survey samples. In survey designs following standard SRS, the selection probability of all units is known before the actual data collection. Hence, weights can be derived as the inverse of the probability of selection. A similar process is followed for ACS, with the probability of selection (within stratum) adjusted to account for the adaptive selection process.

Let n_h denote the total number of starting blocks (PSUs) selected by stratum h , with $h = 1, \dots, H$ indexing each stratum. Let N_h denote the total number of blocks in the universe of stratum h . Within stratum, starting PSUs are selected using SRS without replacement. Whenever the number of informal businesses in a given starting block is above a pre-defined threshold, determined by city, all surrounding blocks are enumerated as long as they are not inaccessible or

⁵ Oruxmaps captures not only the path, but also how long an enumerator stayed in a block, the pace at which s/he is travelling through it, etc.

⁶ The interviews were conducted in Assamese, Gujarati, Hindi, Malayalam, Marathi, Telegu, or English.

⁷ For seminal discussions of adaptive cluster sampling, including issue of sampling weights and proper estimators to use, see Thompson (1990, 1991). Discussions and notation in this section draws heavily, among others, on Thompson (2012), Turk and Borkowski (2005), and Tout (2009).

market centers (market centers are selected by SRS and are not eligible for expansion). The enumeration of surrounding blocks continues until no block meets the pre-defined threshold requirement. This process produces a set of *networks*, which are formed by a group of contiguous blocks that all meet the expansion threshold. Thompson (1990, 1991) defines a network as any group of blocks where the enumeration of one block would lead to the enumeration of all blocks in the network. This definition helps show how, within a network, the probability of selection can be estimated. Specifically, let m_i denote the total number of blocks in a network i .⁸ In the case that a starting block i does not meet the threshold, it is considered as a network with a size of $m_i = 1$. Then, the inclusion probability of a block in network i is defined as π_i and given by⁹:

$$\pi_i = 1 - \prod_{h=1}^H \left[\frac{\binom{N_h - m_{h,i}}{n_h}}{\binom{N_h}{n_h}} \right]$$

The inverse of π_i provides the first-stage weight for each block belonging to network i . For market center blocks, the weight is calculated as $\frac{n_h}{N_h}$.

In providing population estimates using ACS (with stratification), Thompson (1990, 1991) has shown that these weights produce unbiased estimators; however, in many cases there will be informal businesses that are enumerated in blocks that are not part of the starting selection and do not meet the expansion threshold. These blocks are called *edge units*, and the unbiased estimators call for omitting these blocks. The reason for this is that for edge units, there are neighboring, surrounding blocks that have not been enumerated, and so the actual probability of selection for edge units cannot be known. Due to the expense of fieldwork and the fact that there are often interviews obtained in edge units, the choice has been made to calculate weights for edge units as well. These weights are estimated according to the formula for π_i with the following substitutions: the term $m_{h,i}$ is replaced with $(2m_{h,i} + e_h)$ where e_h is a dummy that equals 1 for edge units of strata h and 0 otherwise, and n_h is replaced with $2n_h$.¹⁰ The resulting weights have a known but directionally agnostic bias on population estimates when edge units are retained in this manner. As of the publication of 2022 ESIS in cities in India, we feel it is worth the bias to retain the information collected.

ESIS datasets are also published with variables that identify various groupings of blocks: *M_ID* identifies the network to which a block belongs. *C_ID* denotes what is termed a *cluster*, which is a contiguous group of enumerated blocks, regardless of whether those blocks are from the initial selection, part of expansions, and whether or not they meet the expansion threshold.¹¹ Data users are provided these variables for their own discretion in the treatment of standard errors:

⁸ Note that within a network, all blocks are treated as the same, and so some ACS literature indexes by network rather than block.

⁹ If there are no blocks with stratum h in the network i , then N_h and n_h are set to 0. Note that with networks of size $m_i = 1$, all blocks will be within one stratum and the probability of selection simplifies to the same probability of the initial SRS selection (within the stratum).

¹⁰ In the special scenario where an edge unit is of a stratum that the adjacent network does not contain, the formula uses N_h and n_h of stratum h (i.e., avoiding adjustment detailed in footnote 4 above).

¹¹ It is therefore possible for a cluster to contain multiple networks.

if a user believes that they need to adjust standard errors to account for correlation within a given area, for example, they can choose to apply a clustering option by *C_ID*.

The published ESIS datasets consist of all interviews, respondents of which are selected randomly within blocks. These datasets therefore include second-stage weights that are applied to these interviews. These weights are produced by multiplying the first-stage (block-level) weight by an adjustment factor, which is given by the inverse of the ratio of the number of interviews completed to the total number of informal businesses found in the block. Since within blocks, there typically are both refusals to the enumeration exercise and informal businesses that cannot be reached (and are recorded by observation), assumptions may be needed to calculate this adjustment factor. Specifically, three versions of the second-stage weights are provided based on different assumptions that are used in estimating the total number of informal businesses found in a block, with varying ways of handling refusals and unavailable businesses, as follows:

Assumption	Variable Name	Condition of Inclusion
Strict	wstrict	All confirmed informal businesses only
Median	wmedian	All confirmed informal businesses and refusals that don't have signage/permits on display
Weak	wweak	All confirmed informal businesses and all refusals

The strict assumption calculates the total number of informal businesses found in a block by including only businesses that are confirmed to be informal. The median assumption assumes that businesses that refused to participate in the survey and do not have signage and permits on display are part of the informal sector of the economy; these businesses are added to those confirmed to be informal. The weak assumption treats all refusals, regardless of signage and permits, as part of the informal economy. Therefore, by definition:

$$w_{weak} \geq w_{median} \geq w_{strict}$$

Data users can choose to apply whichever weight they conclude is appropriate under these assumptions. All indicators and analysis conducted by the Enterprise Survey team use the wmedian weights.

Data users should note that there is a debate on whether and how to use weights in regressions (see Deaton, 1997, pp.67; Haider et al. 2013; Lohr, 1999, chapter 11, Cochran, 1977, pp.150). There is no strong, large-sample econometric argument in favor of using weighted estimation for a common population coefficient if the underlying model varies per stratum (stratum-specific coefficient): both simple OLS and weighted OLS are inconsistent under regular conditions. However, weighted OLS has the advantage of providing an estimate that is independent of the sample design. More generally, if the regressions are descriptive of the population then weights should be used. If the models are developed as structural relationships or behavioral models that may vary for different parts of the population, then there is no reason to use weights.

VII. Caveats

Although all possible efforts were exerted to successfully implement this methodology, some informal business units are bound to be missed during enumeration, particularly the type of activities that are hidden on purpose. This is more likely to be the case for household-based activities, although the enumeration process involved, to the extent possible, knocking on every house in each enumerated block to check for informal business activities. Additionally, as explained above, the survey is representative only of informal businesses in the covered geographic area and not necessarily of the entire province or country.

References

- Aberra, A., Aga, G., Jolevski, F., and Karalashvili, N. (2022). “Understanding Informality: Comprehensive Business-level Data and Descriptive Findings”. Forthcoming in Working Paper Series, Washington, D.C. : World Bank Group.
- Aga, G., Francis D.C., Wild M. (2018) “Surveying Informal Enterprises: Applying Stratified Adaptive Cluster Sampling using CAPI with Implementation and Monitoring Tools”, Draft Mimeo
- Aga, G., Francis, D., Jolevski, F., Rodriguez Meza, J., and Wimpey, J. (2022). “Surveying Informal Businesses: Methodology and Applications”. Policy Research working paper, no. WPS 9905 Washington, D.C. : World Bank Group.
- Cochran, William G., Sampling Techniques, 1977.
- Deaton, A. (1997) The analysis of household surveys: A Microeconomic approach to development policy, Johns Hopkins University Press, Baltimore, MD.
- Greig-Smith, P. (1964) Quantitative plant ecology. (2nd ed.) Butterworths, London.
- Haider, S., Solon, G., and Wooldridge, G. (2013) “What Are We waiting for?”, NBER Working Paper 18859.
- Jolevski, F., Aga, G. (2019) “Shedding light on the informal economy: A different methodology and new data.” Let’s Talk Development.
- Levy, Paul S. and Stanley Lemeshow, Sampling of Populations: Methods and Applications, 1999.
- Lohr, Sharon L. Sampling: Design and Techniques, 1999.
- Scheaffer, Richard L.; Mendenhall, W.; Lyman, R., Elementary Survey Sampling, Fifth Edition, 1996.
- Thompson, S. K. (1990). Adaptive cluster sampling. Journal of the American Statistical Association, 85(412), 1050-1059.
- Thompson, S. K. (1991). Stratified adaptive cluster sampling. Biometrika, 78(2), 389-397.

Appendix A: Primary Sampling Units, Sampling Frames

Figure A-1: Hyderabad, Telangana



Figure A-2: Jaipur, Rajasthan

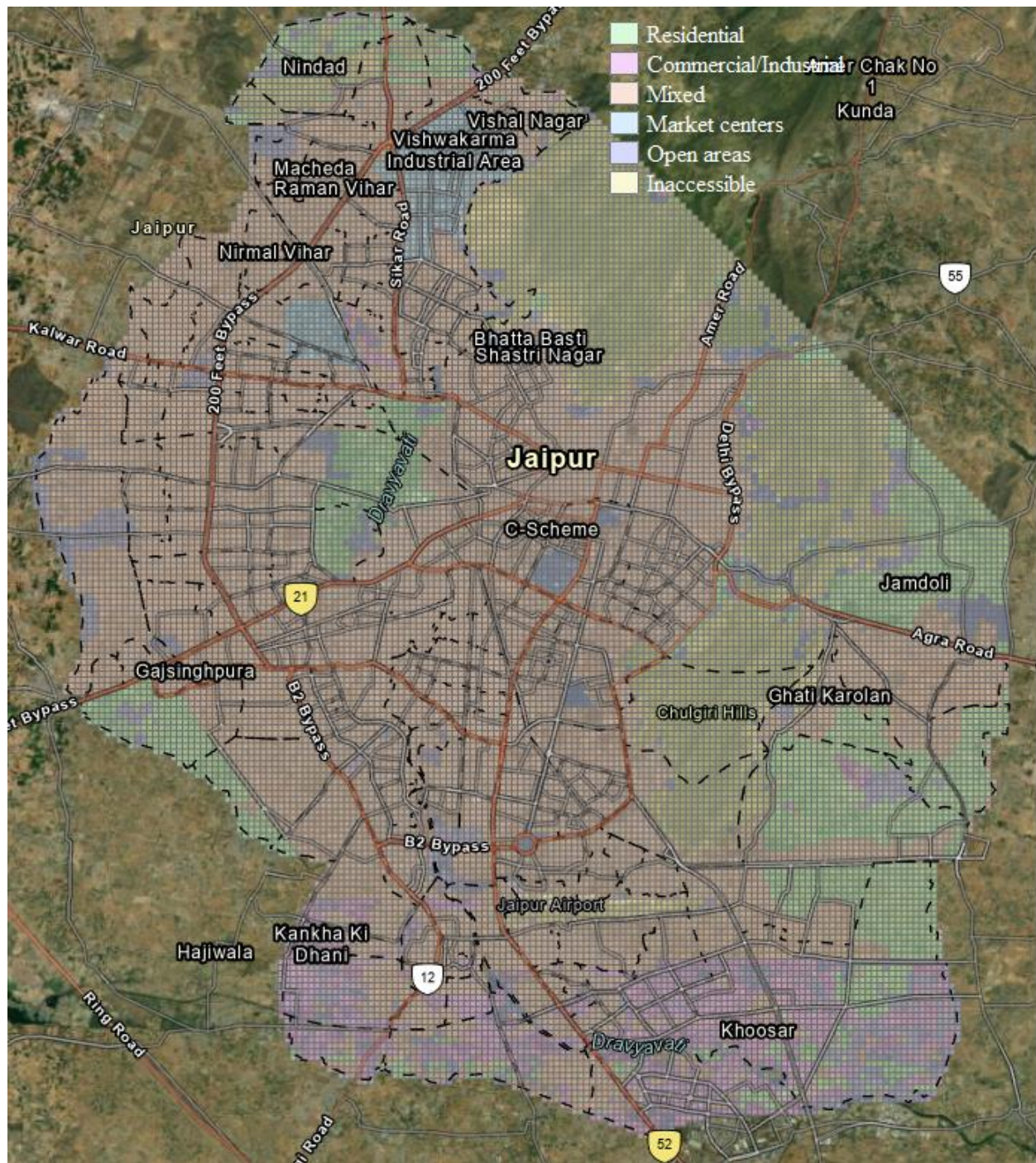


Figure A-3: Kochi, Kerala

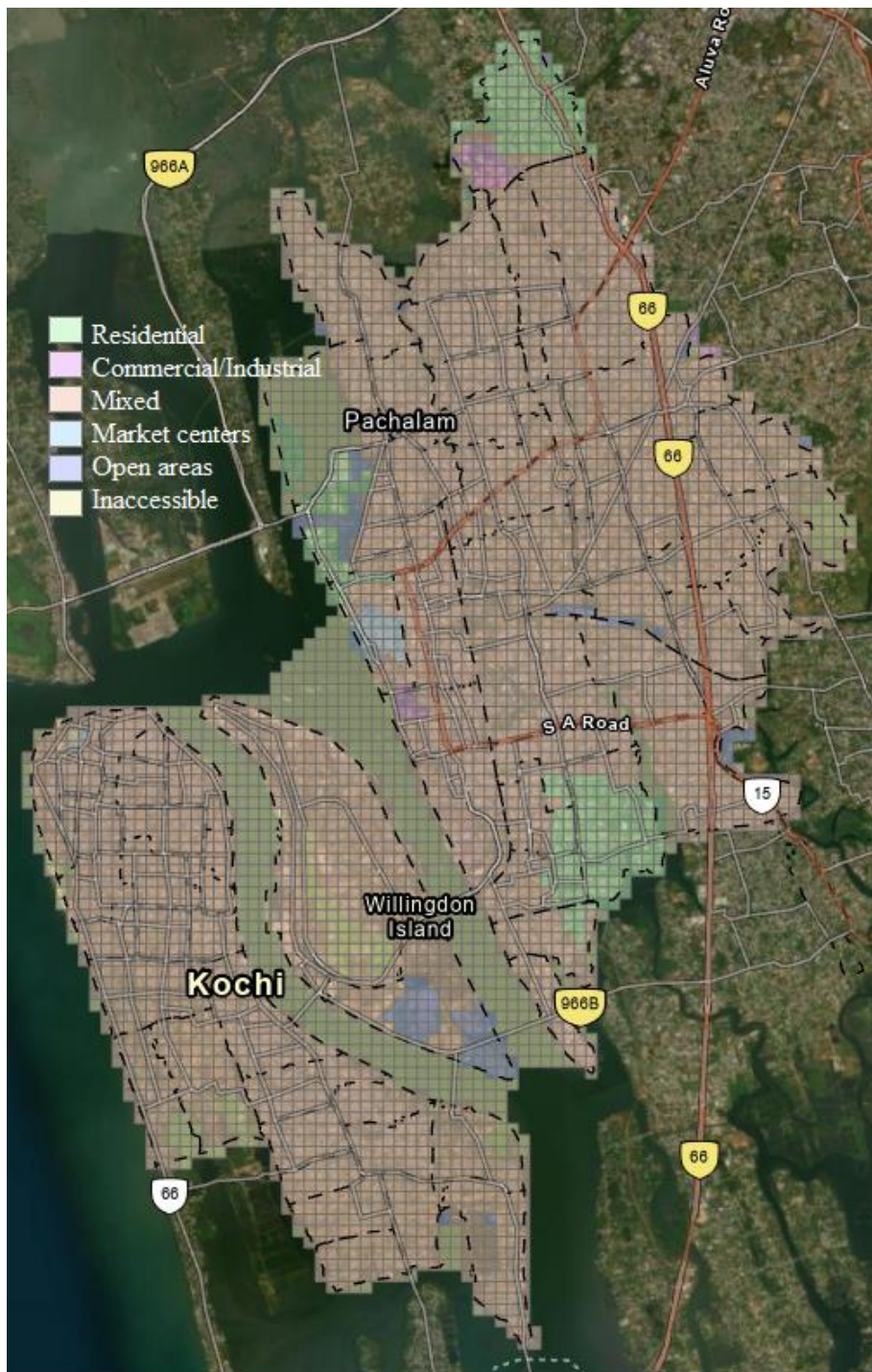


Figure A-4: Ludhiana, Punjab

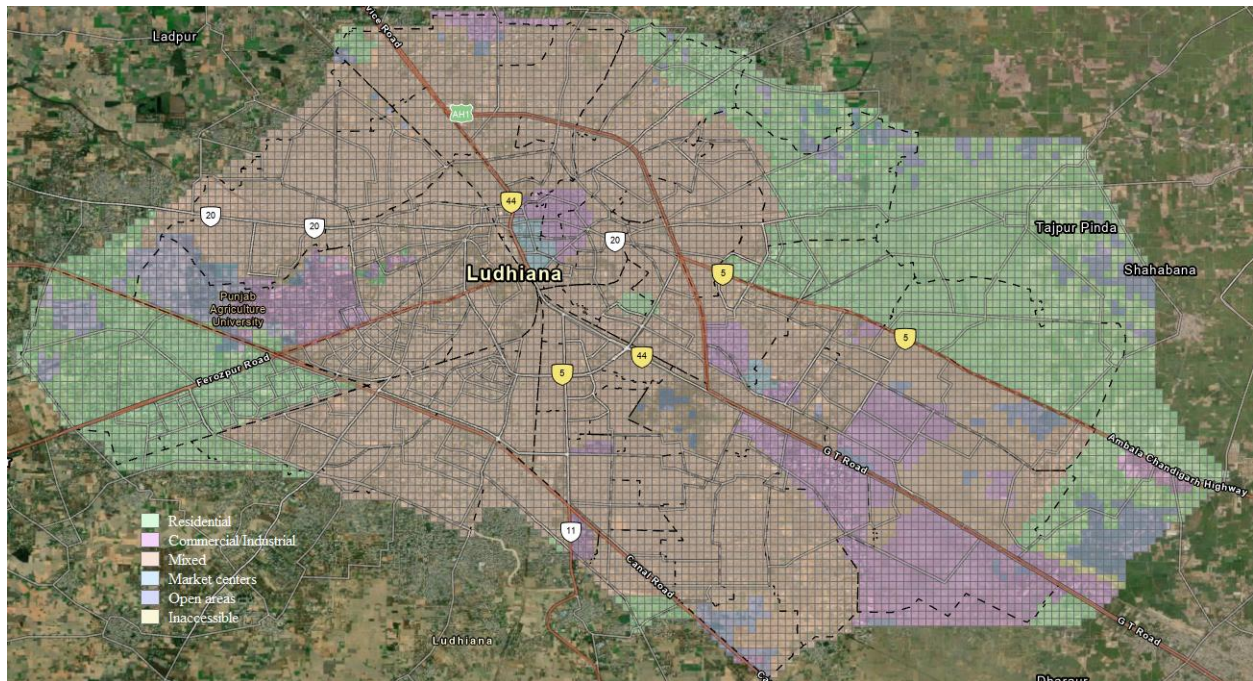


Figure A-5: Mumbai, Maharashtra

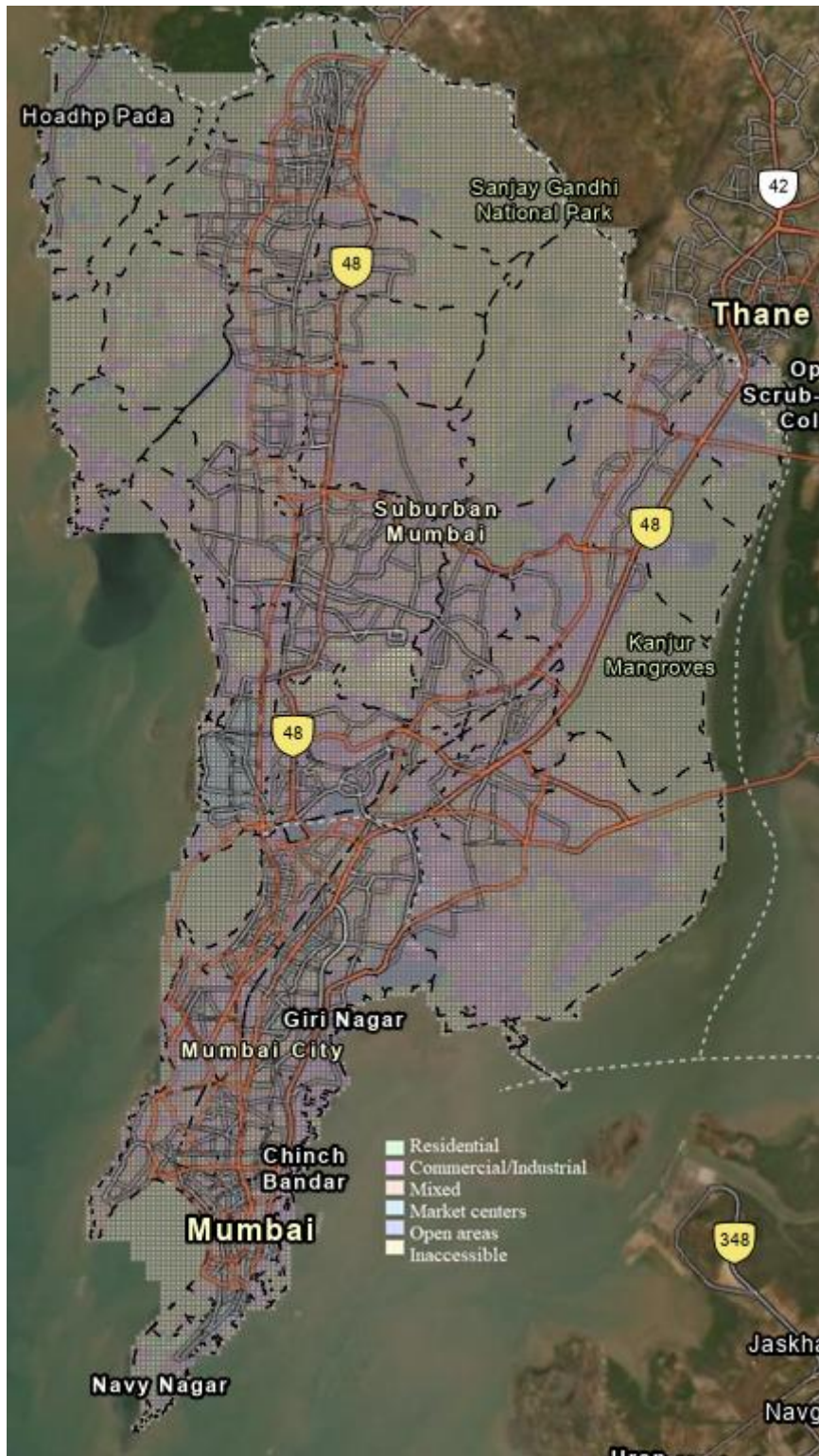


Figure A-6: Sehore, Madhya Pradesh

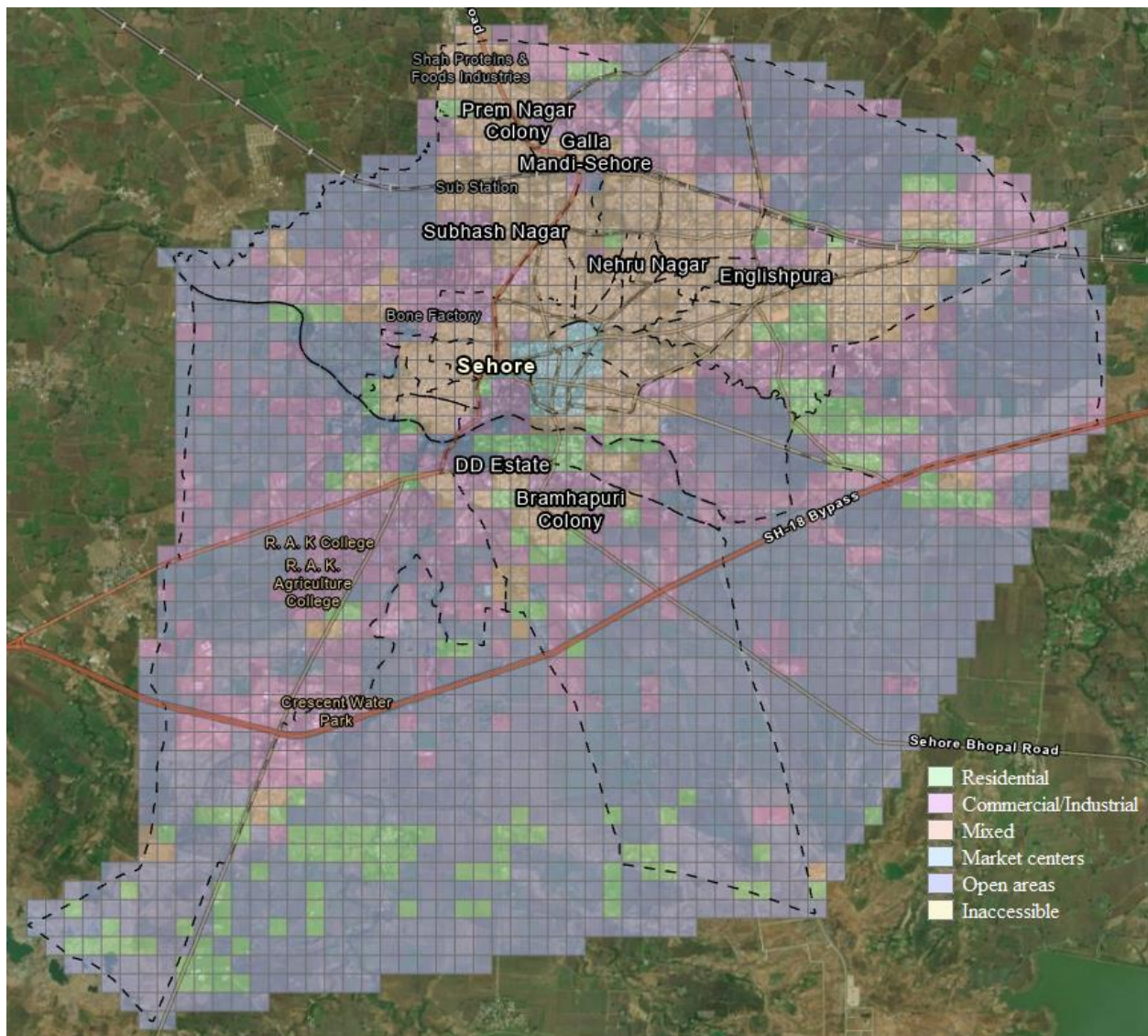


Figure A-7: Surat, Gujarat

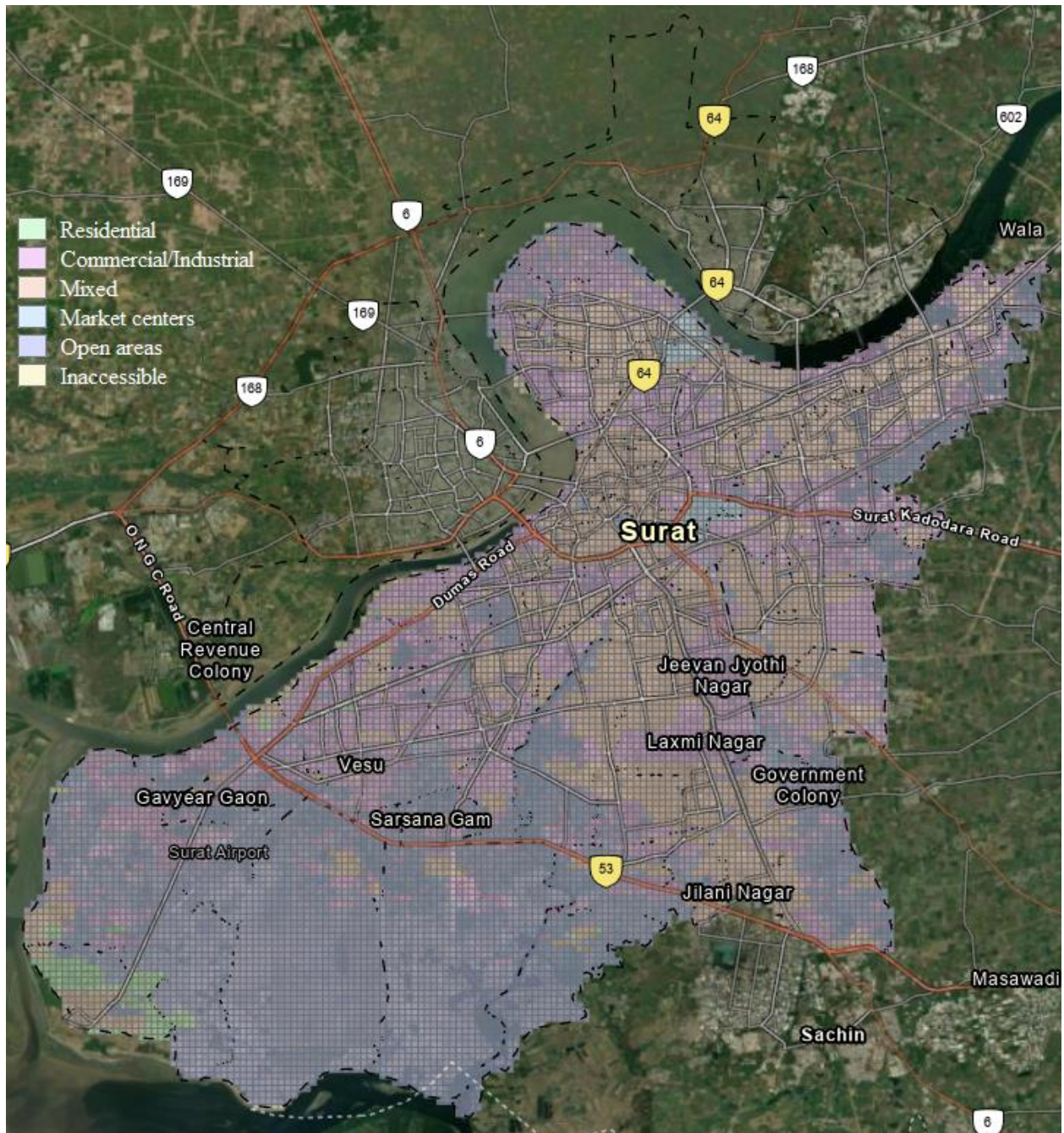


Figure A-8: Tezpur, Assam

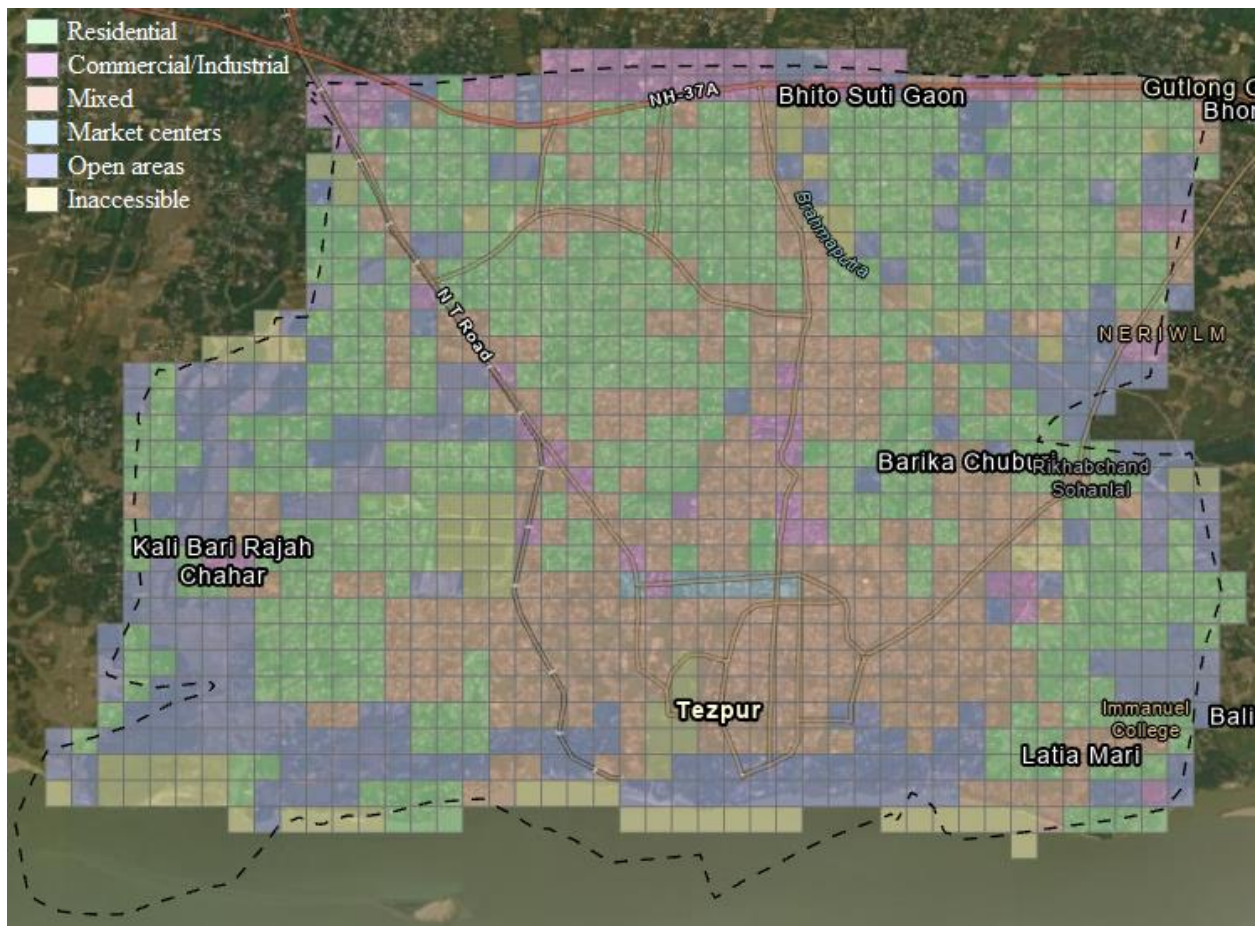
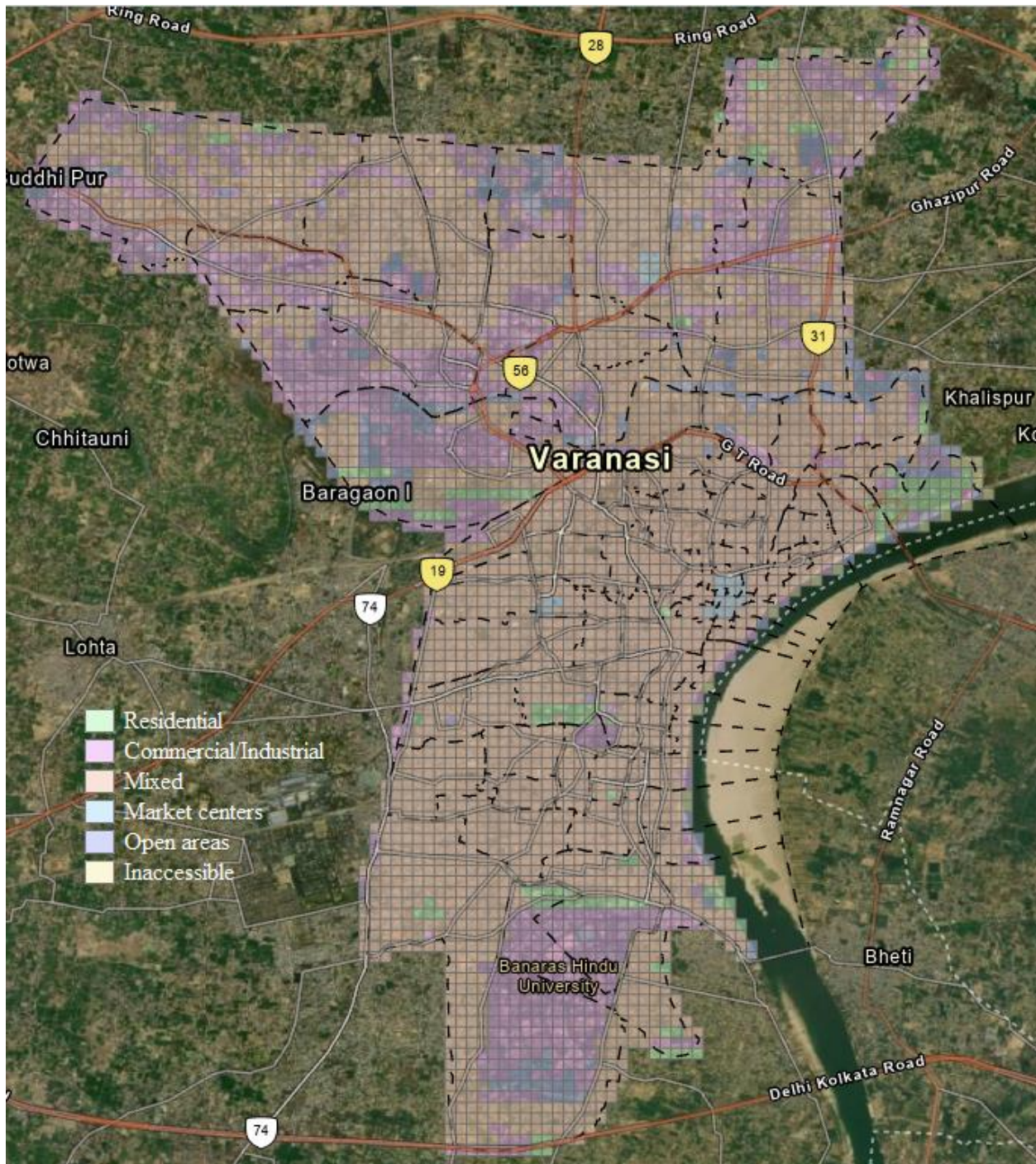


Figure A-9: Varanasi, Uttar Pradesh



Appendix B: Details of Survey Implementation

Table B-1: Number of blocks enumerated, and interviews completed

City	Strata	Universe of blocks	Starting blocks enumerated	Total number of blocks enumerated	Informal business units enumerated	Average number of informal business units per block	Interviews completed
Hyderabad	Residential	1,770	250	281	375	1.33	170
	Commercial/Industrial	1,036	270	280	168	0.60	96
	Mixed	24,146	700	1,283	5,269	4.11	1,558
	Market centers	99	99	99	581	5.87	348
	Open areas	2,257	100	102	21	0.21	19
	TOTAL	29,308	1,419	2,045	6,414		2,191
Jaipur	Residential	2,535	190	205	284	1.39	55
	Commercial/Industrial	1,445	260	288	619	2.15	192
	Mixed	11,670	400	687	2,787	4.06	951
	Market centers	462	286	286	475	1.66	320
	Open areas	1,457	140	154	147	0.95	8
	TOTAL	17,569	1,276	1,620	4,312		1,526
Kochi	Residential	249	40	40	18	0.45	15
	Commercial/Industrial	26	26	26	27	1.04	23
	Mixed	3,243	320	328	817	2.49	518
	Market centers	22	22	22	341	15.50	223
	Open areas	82	8	8	10	1.25	0
	TOTAL	3,622	416	424	1,213		779
Ludhiana	Residential	2,686	130	134	216	1.61	33
	Commercial/Industrial	965	110	118	191	1.62	31
	Mixed	5,040	215	285	1,585	5.56	632
	Market centers	42	42	42	77	1.83	12

City	Strata	Universe of blocks	Starting blocks enumerated	Total number of blocks enumerated	Informal business units enumerated	Average number of informal business units per block	Interviews completed
	Open areas	563	78	79	22	0.28	4
	TOTAL	9,296	575	658	2,091		712
Mumbai	Residential	1,075	77	175	216	1.23	74
	Commercial/Industrial	6,013	430	930	2,017	2.17	561
	Mixed	7,509	750	1,839	5,663	3.08	1,313
	Market centers	529	405	405	468	1.16	397
	Open areas	1,051	38	51	36	0.71	13
	TOTAL	16,177	1,700	3,400	8,400		2,358
Sehore	Residential	163	43	44	22	0.50	7
	Commercial/Industrial	483	68	69	40	0.58	11
	Mixed	305	142	149	668	4.48	371
	Market centers	19	19	19	298	15.68	146
	Open areas	1,130	33	33	4	0.12	1
	TOTAL	2,100	305	314	1,032		536
Surat	Residential	133	80	80	31	0.39	2
	Commercial/Industrial	3,297	372	510	1,379	2.70	273
	Mixed	3,091	554	935	4,446	4.76	1,005
	Market centers	114	114	114	331	2.90	122
	Open areas	4,417	194	215	147	0.68	25
	TOTAL	11,052	1,314	1,854	6,334		1,427
Tezpur	Residential	440	83	118	226	1.92	123
	Commercial/Industrial	63	63	63	144	2.29	63
	Mixed	314	80	113	177	1.57	133
	Market centers	6	6	6	16	2.67	12
	Open areas	235	18	36	48	1.33	26

City	Strata	Universe of blocks	Starting blocks enumerated	Total number of blocks enumerated	Informal business units enumerated	Average number of informal business units per block	Interviews completed
	TOTAL	1,058	250	336	611		357
Varanasi	Residential	139	45	46	70	1.52	25
	Commercial/Industrial	982	95	96	94	0.98	37
	Mixed	2,749	105	144	824	5.72	490
	Market centers	26	26	26	350	13.46	231
	Open areas	217	40	40	38	0.95	3
	TOTAL	4,113	311	352	1,376		786
	GRAND TOTAL	94,295	7,566	11,003	31,783		10,672

Table B-2: ACS parameters, networks, and clusters

City	Threshold for ACS expansions	Number of networks enumerated	Number of clusters enumerated
Hyderabad	11	1,413	892
Jaipur	14	1,274	695
Kochi	45	416	243
Ludhiana	20	575	390
Mumbai	5	1,691	726
Sehore	27	304	93
Surat	11	1,307	532
Tezpur	5	249	48
Varanasi	20	311	204