TPI Technology for the People Initiative

Working Paper

Jumpstarting Spatial Datasets

Sohaib Ahmad Khan, PhD (TPI and LUMS)
Syed Fahad Sultan (TPI)
Muhammad Khurram Amin (TPI)
Hamza Humayun (TPI)
Taimur Ahmed Farooq (TPI)
Marium Afzal (TPI)

December 30, 2014

Technology for People Initiative is based at LUMS, and is focused on helping government departments leverage technology better for improved provision of services to citizens.

The World Bank funded this report. Any opinions expressed in this document are of the authors, and do not necessarily represent the institutional position of Lahore University of Management Sciences or the World Bank.

Lahore University of Management Sciences
Sector U, DHA
Lahore 54792, Pakistan

Table of Contents

Dearth of Spatial Data

Spatial data and its visualizations are absolutely critical in enabling good planning and service delivery for national, sub-national and local governments. A well-designed visualization, which highlights the relevant inferences from a dataset, can make the appropriate decision obvious, and can lead to new insights about the community. It can be used to make a persuasive case for a much-needed intervention to the political leadership, or to rally public support for a new project. Spatial data visualizations can also be used to combat unfair pressure from vested interests, such as in the case of deciding the location of a new school or health facility, by enabling greater transparency. Indeed, GIS departments are now the mainstay of modern city and subnational governments, and it is difficult to imagine effective service delivery or resource provisioning without them.

Yet, many developing countries suffer from an acute shortage of appropriate spatial datasets and the capacity to utilize them in their daily affairs. The dearth of spatial datasets that can enumerate and help visualize the socio-economic well being of a community, is a serious impediment to governance, service-delivery and development-planning in most third-world countries. As a result, decision-making is often 'blind' and informed on hearsay or whims, rather than actual evidence.

Why is *spatial* data particularly important? This is primarily because most human activities are linked directly or indirectly to space. Governance units, such as city neighborhoods, police precincts, school districts, rural villages, canal distributaries, farmlands and tax circles, are all fundamentally spatial entities. Facilities such as hospitals or schools service a certain spatial area, and population is also spread in spatial clusters.

Moreover, humans perceive images and visualizations much more effectively than data stored in tables or registers. Well-designed spatial visualizations have the capability to show a large amount of data in a palatable form, which highlights its primary trends and reveals relationships. On the contrary, such relationships are not easy to infer from table-based data.

In Pakistan, most government offices operate without access to spatial datasets and visualizations that are essential for their work. The dominant data collection and processing practice in government offices is to employ a set of handwritten *registers*, which have ruled pages, with columns storing different attributes of a data record. The record of patients in a hospital, or the record of criminals in a police station is kept in similar registers. Many government departments have a notified set of official registers – for example, a police station is required to keep a set of 25 registers, a practice dating back to the British era. Many of these registers are essentially different queries on the same dataset. For example, one of these registers may be indexed by the investigation officer assigned to a criminal case, and another may be indexed by proclaimed offenders that have perpetuated these crimes. The register is the quintessential data recorder at many levels of government.

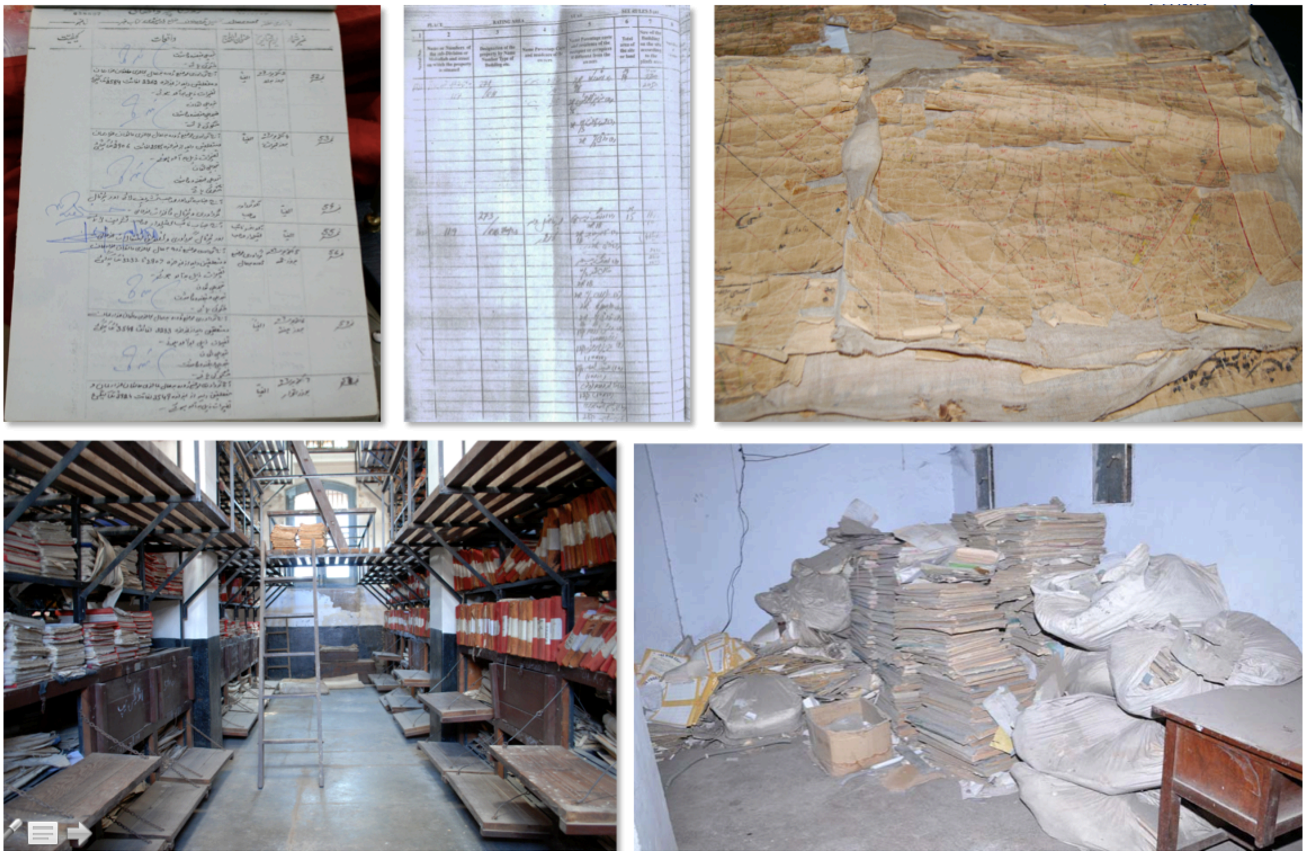


Figure 1: Few examples of spatial data records: *Top left:* A page from patwari's *basta*, land revenue record. *Top center:* Register to keep record of properties in property tax administration. *Top right:* A damaged *masavi*, cadastral map of village. *Bottom left:* District record room, which keeps record of land revenue, court cases and DC's office, including valuable spatial data. *Bottom right:* Record room of a court, containing case files.

Many of these registers contain valuable spatial data. For example, the FIR in a police station lists the location where the crime has taken place, and subsequent investigative reports (*zimni*) attached to the case record may even contain a sketch map of the crime scene, prepared by the investigation officer. Yet this spatial data is hardly integrated at a higher level, such as to develop an understanding of the underlying crime trends and their response. Similarly, school districts will not have a spatial understanding of the 'catchment' areas that their schools serve, or an understanding of communities in their jurisdiction that are underserved.

What is required to build such an understanding? To take the example of a school district further, consider the problem of a government officer in charge of a school district (EDO Education) having to decide on the location of 5 new primary schools that he has the funds for. There are several pieces of data that are relevant to make this decision. The EDO should ideally have information about:

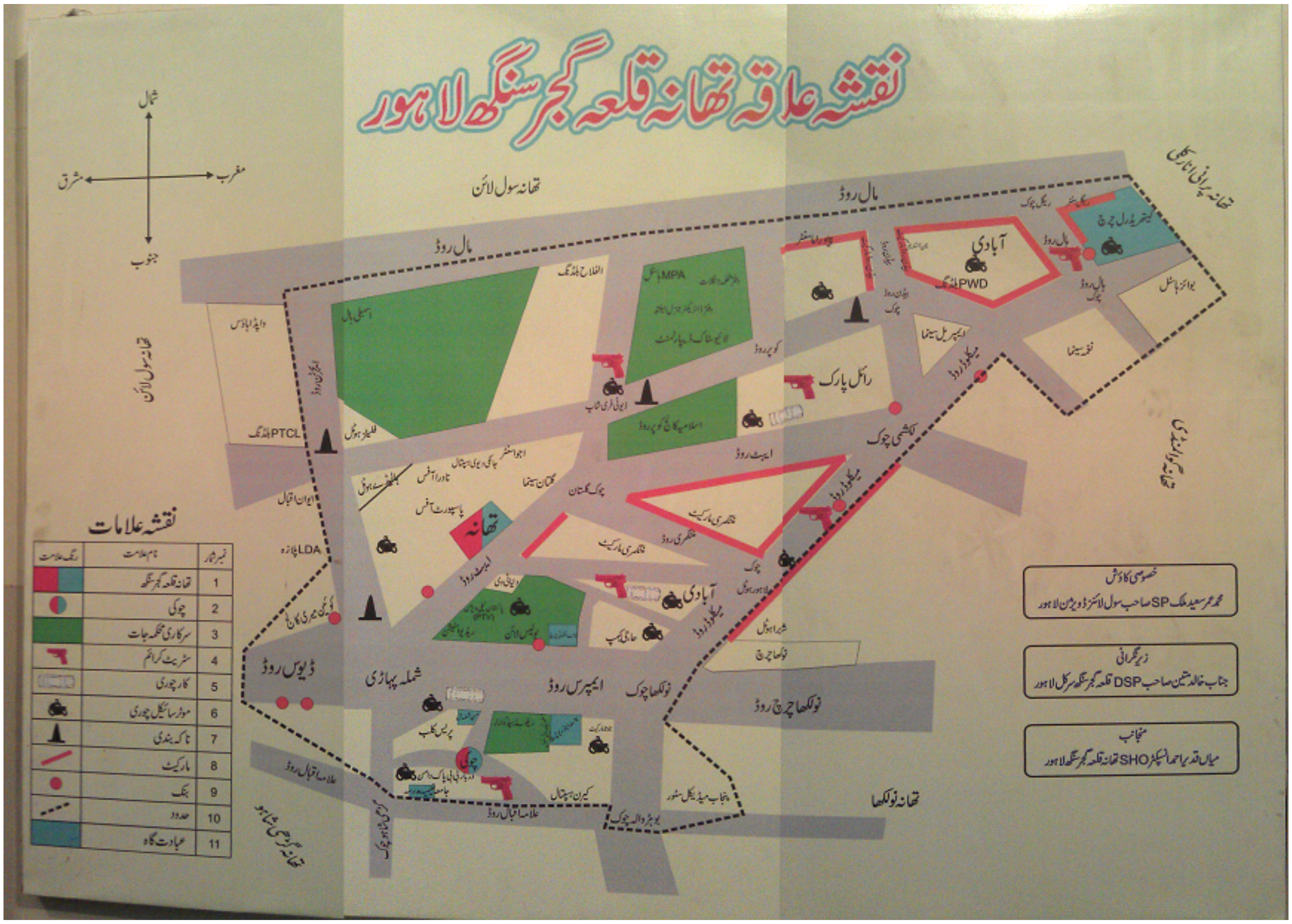
- Population: What is the density and spread of people in the district.
- Village locations
- Existing literacy information of each village, and information about number of school-age children
- Locations and capacity of existing schools.
- Addresses of students in existing schools, to determine how far students have to travel to reach their school.
- Addresses of teachers, to determine their travel time and availability in an area.
- Road network layout, to determine accessibility.
- Terrain information, to determine accessibility

Note how each aspect of this data requirement has a strong spatial component. If the EDO can visualize where the existing schools are (supply) and where the potential students are (demand), and can also determine how much distance and time students are spending to reach existing schools (accessibility), then determining the location of new schools can be done in a fair and transparent manner.

We will use this problem of locating a new school as a running example in this report for illustration. Many other problems are similar in nature as far as data needs are concerned, where they are locating new infrastructure, monitoring service delivery, or measuring development indicators. All require an understanding of the spatial spread of population and the services they are delivered or their socio-economic well-being.

Impediments in Using Spatial Data

While mapping is an age-old activity, GIS systems and equipment has recently become mainstream and affordable. However, when we look at the landscape of spatial information in Pakistan, we find a lot of basic ingredients missing, which lead to poor adoption of spatial technologies. In visits to several local and sub-national government offices, including departments of police, education, property tax, land-revenue, judiciary, health and disaster management, we did not find adequate adoption of data-driven decision making. In fact, whatever spatial visualizations were available were more for aesthetic purposes and for adornment of walls rather than for any data-driven decision-making. In one police station, we found an age-old map shown in Figure 2, which is hardly usable and is infested with major errors, including distortions of scale and incorrect determination of ordinal directions. Interestingly, when we developed an up-to-date map of that police station and presented it to them, police officials from other units who had no reasonable use for this map requested a copy to decorate their office walls!



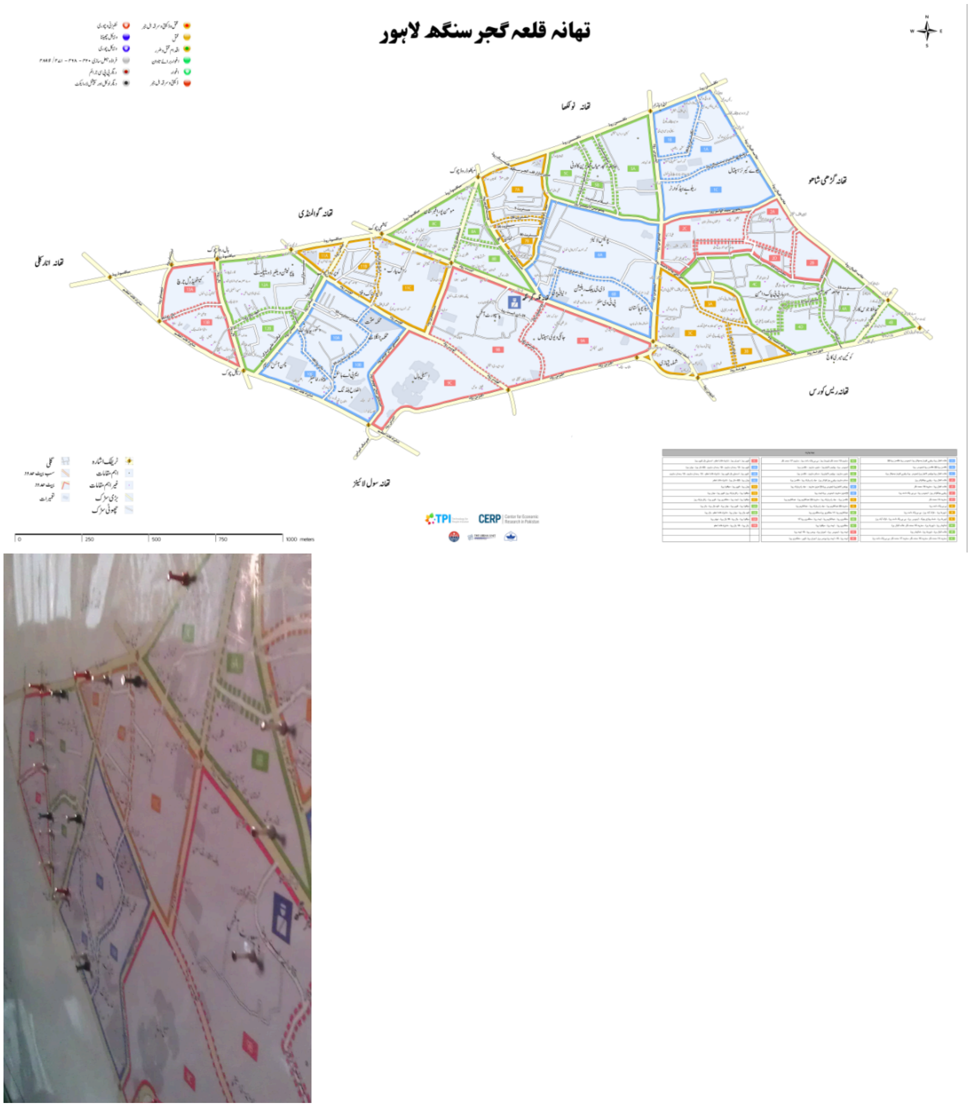

تھانہ قلعہ گجر سنگھ لاہور
N
W
E
S
0
250
500
750
1000 meters
TPI
CERP Center for Economic Research in Pakistan

Figure 2: *Top:* Map of the jurisdiction of a police station, found on the wall of its Muharrar's office. The map's legend of ordinal directions has East and West directions flipped, while the actual map is flipped in North-South direction. The scale is distorted and it contains static hot-spots. *Bottom:* An up-to-scale map prepared by our team, with beat boundaries. This map was mounted on a soft-board and police staff was encouraged to use pins of different colors to show updated crime trends.

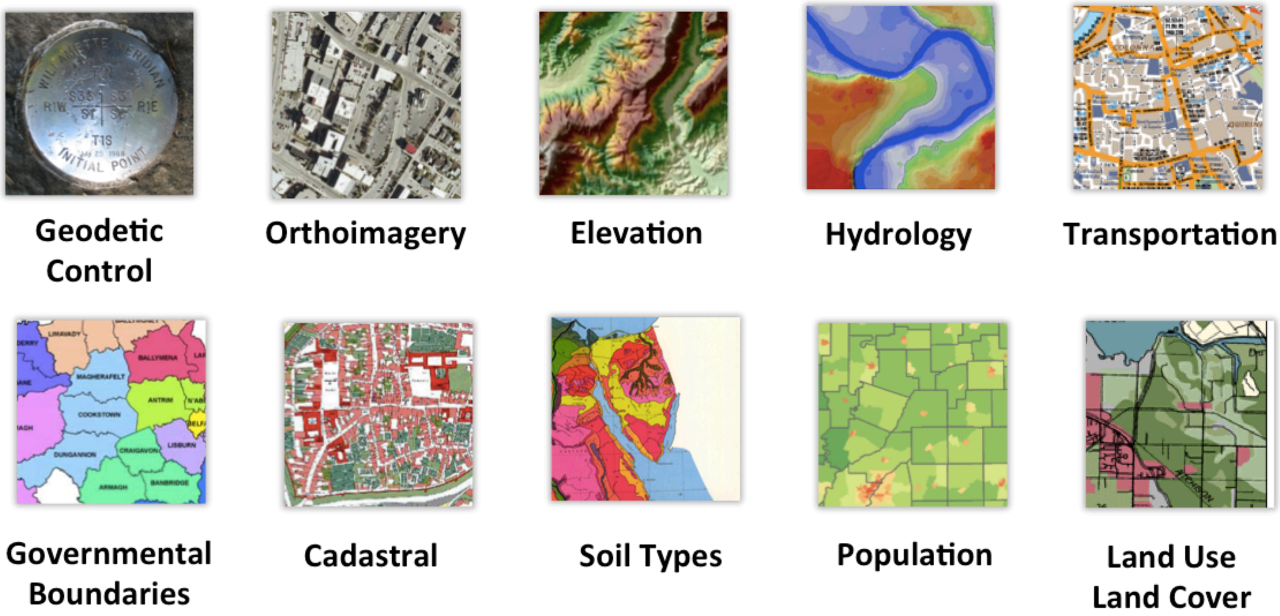


Figure 3: Base layers of spatial data that are often needed for common data analysis tasks. Of these, elevation, hydrology, transportation, governmental boundaries, population and land-use-land-cover are extremely critical in our opinion. Non-availability of these layers hinders adoption of spatial data and results in duplication of effort because every project needs to generate these.

Lack of Availability of Base Layers

It is our observation that the critical missing ingredient in adoption of a data-driven approach is not the technology or the availability of software and tools, but rather, the non-availability of basic datasets that are needed for almost every spatial analysis that is needed for governance. In the theory of GIS systems, it is known that certain fundamental datasets are needed for most practical projects. These are called base-layers in the terminology of Spatial Data Infrastructures (SDI). It is postulated that development, maintenance and dissemination of these base layers should be a coordinated governmental activity, and this data is part of the infrastructure that government provides, pretty much like the infrastructure of roads or sewage. See Figure 3 for some of the critical base layers that need to be development through committed governmental effort in Pakistan.

Even in the absence of standardized base layers, there is still a lot of spatial data that exists but is in a form that is not very useful. Below we highlight some of the critical problems that we have seen, which, if fixed, can encourage the use of spatial data in the country.

Lack of Standardization of Addresses

The most ubiquitous form of spatial data is the address of a citizen or a business. Addresses are recorded on every National ID Card (NIC), school enrollment forms, electricity bills, FIRs, patient data in a hospital or BHU and pretty much everywhere a governmental service is to be delivered. The address refers to a spatial location, and hence, theoretically, the collation of addresses on a GIS map can generate the required spatial visualization for a government department. In our schools example, the address of all students, if plotted on a map, will generate the visualization of catchment area of the school, and can inform the decision maker about whether a new school should be located within that area, say to reduce the travel time of students.

Yet, this problem is deceptively difficult for two fundamental reasons. Firstly, in Pakistan and many developing countries, addresses are not standardized, and can be written in many different forms. This makes it very difficult for a computerized system to automatically process the address and place it on a map. While humans can still manually understand the address, mostly because of their familiarity with the area, automatic processing becomes very difficult. Secondly, even when fields of an address can be understood by a computerized system, the data needed to place it on a map is not available. For example, the address of “Qadir Street No. 1, Jeewan Hanna Kachi Abadi, Garden Town, Lahore” refers to which road segment on the map is not easy to determine. This is because the maps available from Survey of Pakistan, Google Maps or Open Street Maps do not record addresses down to a detailed level. This database of place names, while improving rapidly on crowd-sourced platforms such as Google Maps and Open Street Maps, has a long way to go before it can become usable for an automated system.

An alternate way of addressing which is prevalent in older parts of cities and rural areas is landmark based. That is, instead of defining the address in the hierarchy of house, street and neighborhood, the address is defined in terms of the nearest landmarks, using terms such as ‘opposite’ or ‘near’. Even the address of our university, LUMS, which is a premier educational institution of the country, contains the term ‘Opposite Sector U’, which would not be acceptable in a standardized addressing scheme employed in western countries. After all, if Sector U is a square region, there are four sides on which the ‘opposite’ could be located! In older parts of the city, terms like ‘Near Masjid Allah-hu’ or ‘Mazar Bibi Pakdaman’ may be commonly seen in addresses. Since Mazar (tomb) of Bibi Pakdaman is a point entity on the map, how much area in its proximity is addressed by its reference is not straight-forward to determine. Yet older postal addresses sometimes show just the name of the neighborhood, with the statement addressed to the postman saying “Ahmad Sahib to melay”, literally, “this letter should reach Ahmad Sahib”.

If this is the state of addressing, a natural question to ask is how does government function? How does post get delivered? How do policemen find a location?, How do land transactions happen? How do electricity distribution companies locate consumers who are not paying their bills? The answer to all these questions is quite interesting. Addressing does indeed work, but not in the global sense where the address can be resolved at one location. Instead, resolving an address to a spatial location is a distributed activity, that, because of the absence of detailed maps, cannot be completed at one location. Rather, it works in a distributed fashion in a similar manner to the way packets get routed on the Internet. On the Internet, at each router, only the minimal information about the next hop of the packet is decoded. Yet, by a sequence of these next hops, the packet reaches its destination. In exactly the same way, a letter from Rawalpindi, addressed to say, 17, B-Block, DHA, Lahore, will be put in the ‘Lahore’ bin at the Rawalpindi GPO. When it reaches Lahore GPO, there is a sorter there who puts all incoming mail into one of the 43 bins corresponding to the 43 post offices in Lahore. Therefore this letter will find its way to the DHA bin. At the DHA post office, it will again be sorted to the bin of the postman who is assigned to the B-Block beat. That postman knows where house 17 is.

Interestingly, the precise address is known within the system, that is, the postman who walks the beat understands the layout of the area. The problem, therefore, is not that mail will not get delivered. The problem is that there is no central location within the system where all addresses can be located, and hence, no visualization of, let’s say, the load of mail delivery process in entire Lahore can be generated. The situation is similar in locating villages, or land plots in land revenue department, where the local patwari knows how to locate a Khasra number, but the higher ups cannot find that location on a map unless they seek help from the local patwari.

Lack of Hierarchical Administrative Units

Even when we are unable to locate a precise address or street segment, for many applications, it should be sufficient to locate the administrative region in which that address exists. For example, in rural areas, the NIC contains the mauza (village) name of the citizen. If all these mauzas were available as polygon shape files on a map, then we would be able to locate the citizen within that polygonal region, even if her exact house address was not traceable. This would be sufficient for many socio-economic applications. Similarly in a city, knowing the bounds of a neighborhood, such as Jeewan Hanna Kachi Abadi, or Garden Town, in the above example, should be good enough for many applications. This leads us to a very significant problem – the standardized definition of administrative units, and determination of their boundaries.

This, again, turns out to be a deceptively hard problem that requires dedicated work. In rural areas, the situation is somewhat simpler, because of the well defined hierarchy of the land-revenue system. The mauza, or revenue village, exists within a patwar circle, which is part of a qanoongoi, which combine to form a tehsil, which is in a district. The problem is simply to locate the polygonal boundaries of the smallest unit, the mauza, and then information can be integrated upwards at the desired level of granularity, because every subsequently coarser unit is made from the combination of this smaller unit, which essentially is the building block of this hierarchy. It is also critically important to observe that all other relevant units used in governmental work, such as the police station, electoral constituencies, school districts, or union councils, also are created from the same building blocks. This enables integration of data from different sources of analysis and planning.

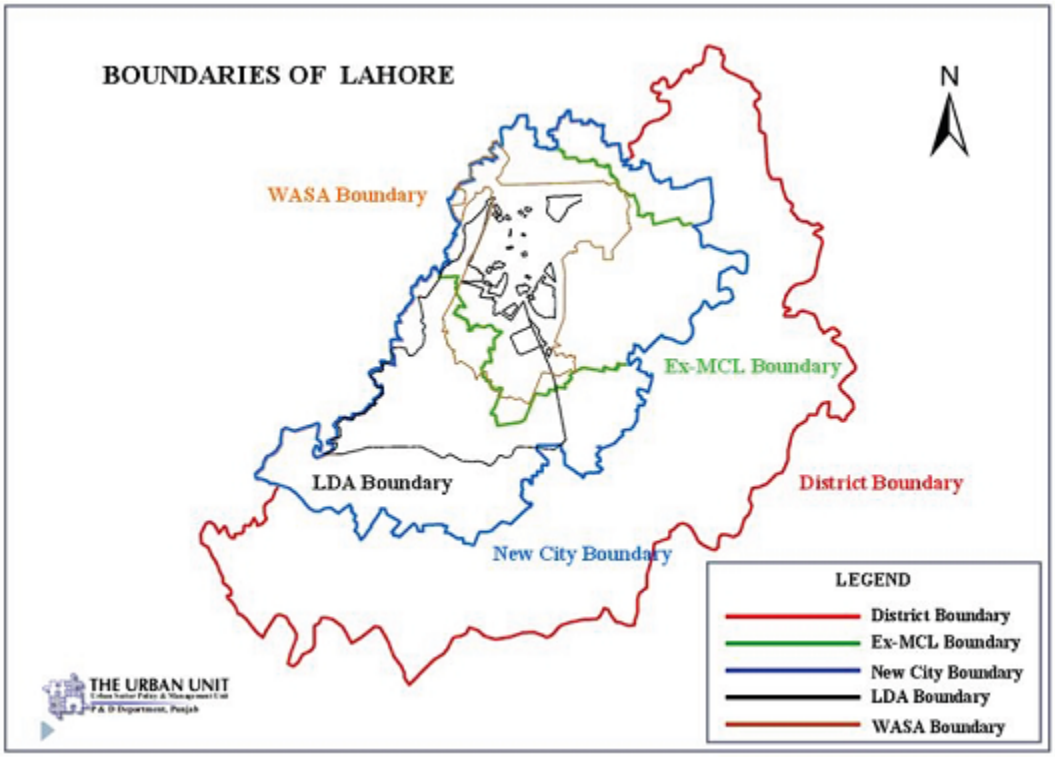


Figure 4: Urban boundaries of Lahore, as reported by five different government departments. Their inconsistency highlights the challenges in integrating data from different sources, which is critical for any socio-economic analysis. [Figure courtesy of the **Urban Unit, Government of the Punjab**]

The scenario is not as clear in urban areas. Administrative units in urban areas are characterized by overlapping and ambiguous jurisdictions, mainly because the administrators themselves have higher visibility of the region under their jurisdiction, and can therefore, choose to modify their units according to their needs. A somewhat stark example of this phenomenon is illustrated in Figure 4, which shows five different boundaries of Lahore collected by the Urban Unit from different governmental agencies. The disagreement between the definition of Lahore itself is extreme, and can lead to a lot of problems in data analysis. For example, with this disagreement, the population of Lahore would be perceived to be grossly different by these agencies.

At a sub-city scale, we compared the polygonal boundaries of Union Councils (important for electoral data and political constituency) with that of police stations

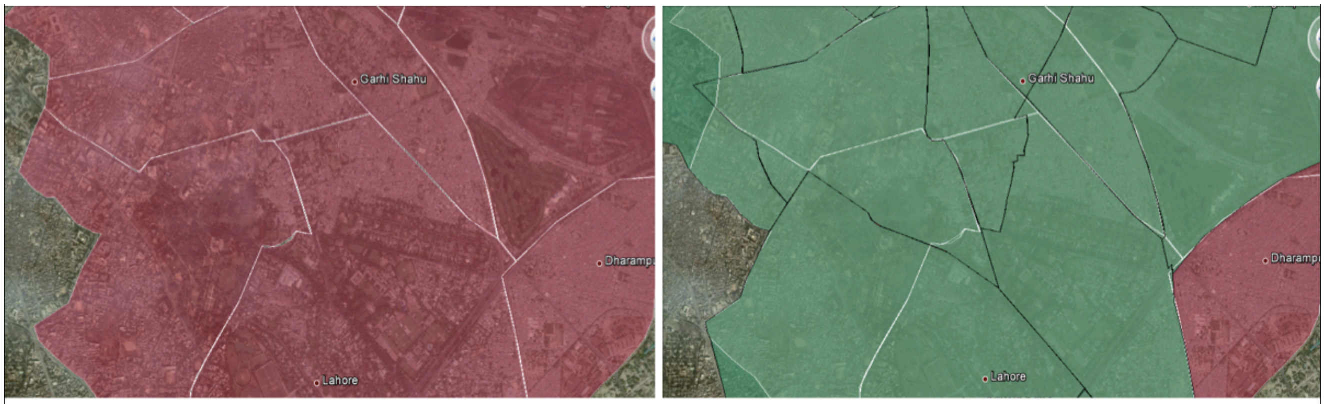
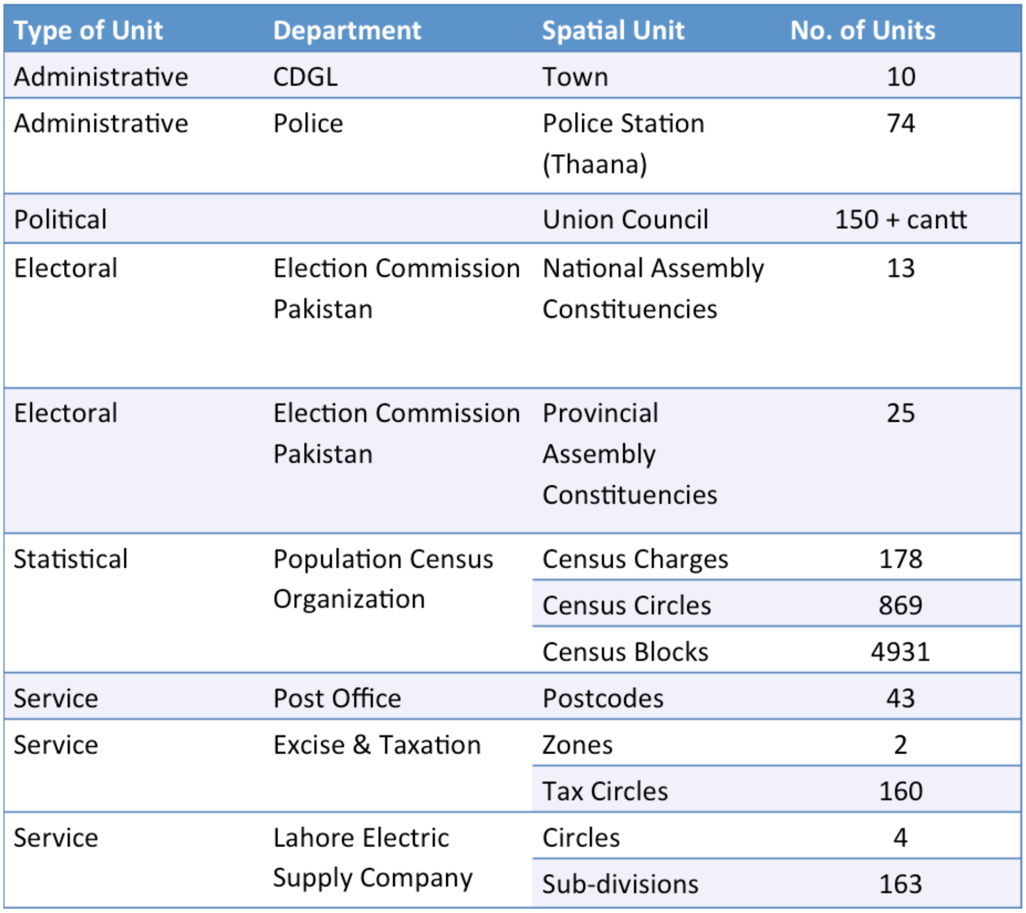


Figure 5. Boundaries of police stations in Lahore (*left*) overlaid with boundaries of union councils (*right*). While consistent in some areas, they are grossly different in others.



Type of Unit	Department	Spatial Unit	No. of Units
Administrative	CDGL	Town	10
Administrative	Police	Police Station (Thaana)	74
Political		Union Council	150 + cantt
Electoral	Election Commission Pakistan	National Assembly Constituencies	13
Electoral	Election Commission Pakistan	Provincial Assembly Constituencies	25
Statistical	Population Census Organization	Census Charges	178
		Census Circles	869
		Census Blocks	4931
Service	Post Office	Postcodes	43
Service	Excise & Taxation	Zones	2
		Tax Circles	160
Service	Lahore Electric Supply Company	Circles	4
		Sub-divisions	163

Figure 6: Table of constituent units of different administrative divisions in Lahore district. These units are not hierarchically aligned, making data integration very challenging.

(Figure 5). With roughly about twice as many police stations as union councils, with a bit of careful planning, the two sets of boundaries could be aligned, that is, the boundary of the bigger unit (in this case, the union council) should be form from aggregation of the smaller units (in this case, the police station). This is particularly reasonable to achieve, especially because both units are ideally constructed as a function of population – if population in one union council or police station exceeds a certain limit, it should be bifurcated into two. Upon comparing their boundaries, we found that at places, they were in agreement, but there were also large differences between them. This seems simply a result of not making effort towards coordination of boundaries, rather than for any other fundamental reason.

Why is this difference an undesirable feature from data analysis point of view? Because it makes integration of data very difficult. For example, we have population estimate of each union council available in census data, and we have crime numbers of each police station available with the police department. If these boundaries were indeed aligned in a hierarchical fashion, it would have been easy to develop a visualization of crime per capita at the union council level. With current arrangement of overlapping boundaries, it is a quite difficult to create this visualization.

Population Estimation

One of the most critical and frequently needed data layers is that of population. Since governmental services are targeted towards citizens, it is imperative to know where the citizens are. In the problems of targeting humanitarian aid after a disaster, planning a new school, or provision of police resources according to crime incidents per capita, it is of utmost importance to have a good population density map available to the decision makers. Yet, despite its fundamental importance, it is one of the layers that is missing from the landscape of spatial data in Pakistan. A good, detailed and up-to-date population layer does not exist, and even coarse approximations are unavailable.

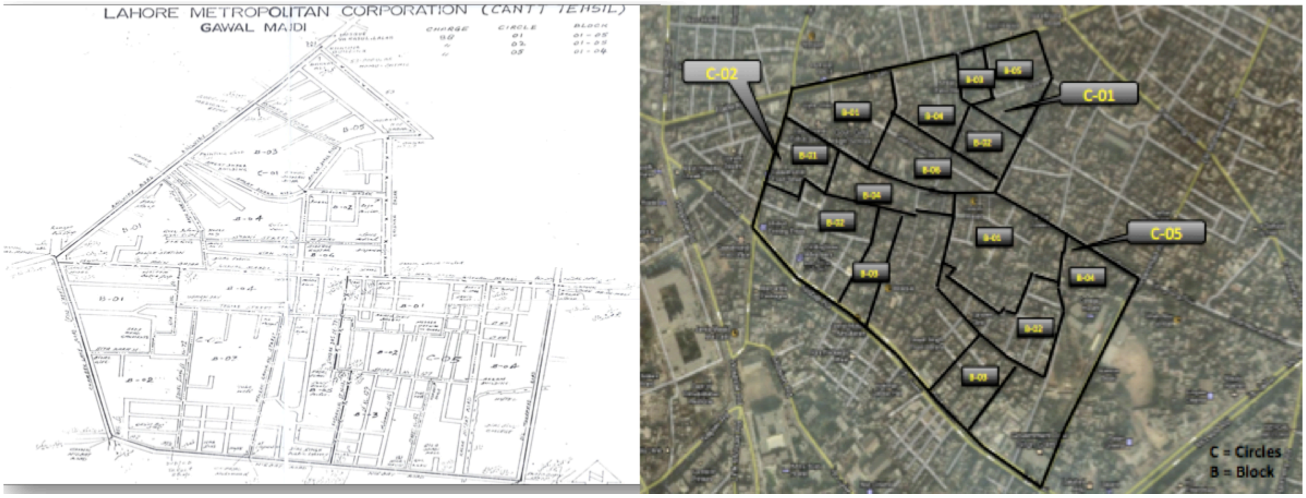


Figure 7. A sketch map of census blocks used by the census department (*left*), and its actual spatial layout (*right*) which is quite different in shape, orientation and scale.

The primary enumeration of population is through the population census, which was last conducted in 1998. Even this dated census data has not been located on a map. The census methodology revolves around enumerating households in a census block, which is typically targeted to have 250 households. Census blocks combine to form a census circle, and these combine to form a census charge. The problem is that these units exist in tabular form, with text descriptions of their demarcation, but are not drawn on any map. Hence, the population data cannot be transferred to a GIS map.

Over the last couple of years, Pakistan Bureau of Statistics has started the effort of demarcating these blocks in spatial coordinates, by using GPS devices in the field and marking every landmark location contained in the text description of the boundary. This is indeed a commendable effort in the right direction, and more resources should be dedicated to completing this layer. This sort of data needs continuous maintenance, because of the updates in these boundaries, and will also need to be made publically available to citizens and government departments, so that its potential benefits can be realized.

In the current scenario, where such detailed population data is not available, the goal of informed decision-making based on spatial visualizations is really very difficult to achieve.

There can be several proxies to population data. For example, even if population is not known, but the size of each settlement is known on a map, it can be used to approximate population. However, that is also a form of data that is unavailable in Pakistan. At a very coarse level, the population estimate of each tehsil can be shown on a map, but for most sub-national governance tasks, a tehsil is too coarse a unit to work with, and it can mis-inform decision making by hiding the non-homogeneity of socio-economic conditions within such a large area.

An associated problem, though not one of spatial visualization, is that of the accuracy of census population data itself. The last census was conducted in 1998, and since then, the population estimates are generated by compounding a fixed percentage of growth every year, uniformly across the whole country. This is hardly an accurate strategy, and does not take into account the relative movements within the country, such as migration to cities, or movement trends between cities. Over the course of more than fifteen years, one would expect such trends to shift, and hence, an update of the census data becomes critical. One way to achieve this, in the absence of a full-fledged census, can be to conduct small sample surveys in randomly selected census blocks, and use their data to adjust the cumulative percentages that update population estimates.

Opportunities to Innovatively Jumpstart Spatial Datasets

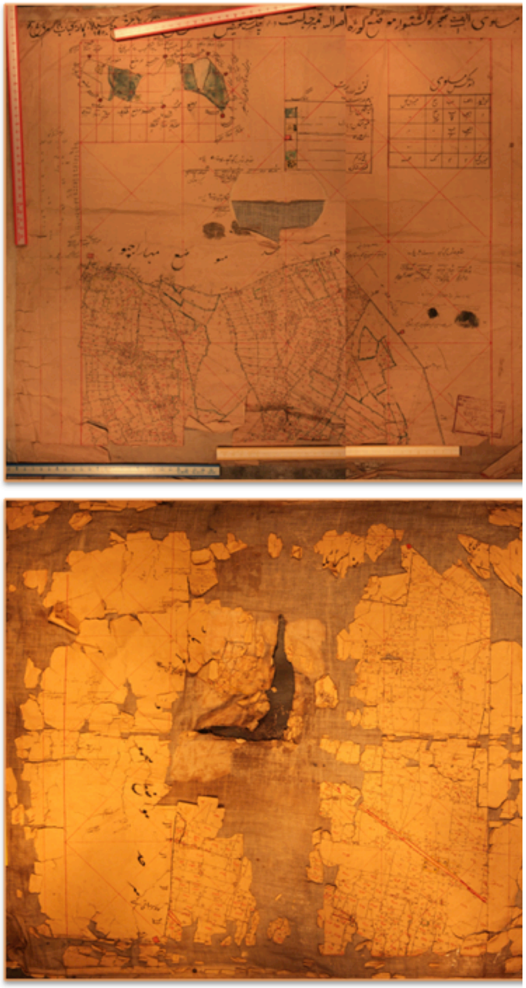
While the landscape of spatial data is not rosy, what is encouraging is that in recent years, many new techniques have been developed which hold great promise. The players in the field are changing. While previously, mapping used to be the domain of one central and well-funded agency, now most consumer cell-phones contain GPS which is a powerful mapping device. Hence, the landscape is now characterized by many actors rather than a single agency, and data is being made available from a multitude of sources. Satellite imagery, Open Street Maps (OSM), and powerful computers can all be utilized to jumpstart spatial datasets and visualizaitons, without the need for expensive field surveys in many cases. In the following sections, we highlight several such attempts which generate spatial data at a mass scale, often of less than perfect accuracy, but still extremely useful in a scenario where none was available before.

Developing a Mauza-Level Administrative Boundary Map for District Jhelum and its Use in Socio-Economic and Demographic Mapping

A Step Towards Building a National Spatial Data Infrastructure for Pakistan

Background: Our team consisting of experts from SUPARCO and LUMS developed an *innovative, scalable and cost-effective methodology* to generate an accurate mauza(village)-level boundary map of Tehsil Jhelum. The work on this project was completed in a record time of less than 8 weeks, in partnership with the Population Census Organization and the office of the District Officer Revenue, Jhelum.

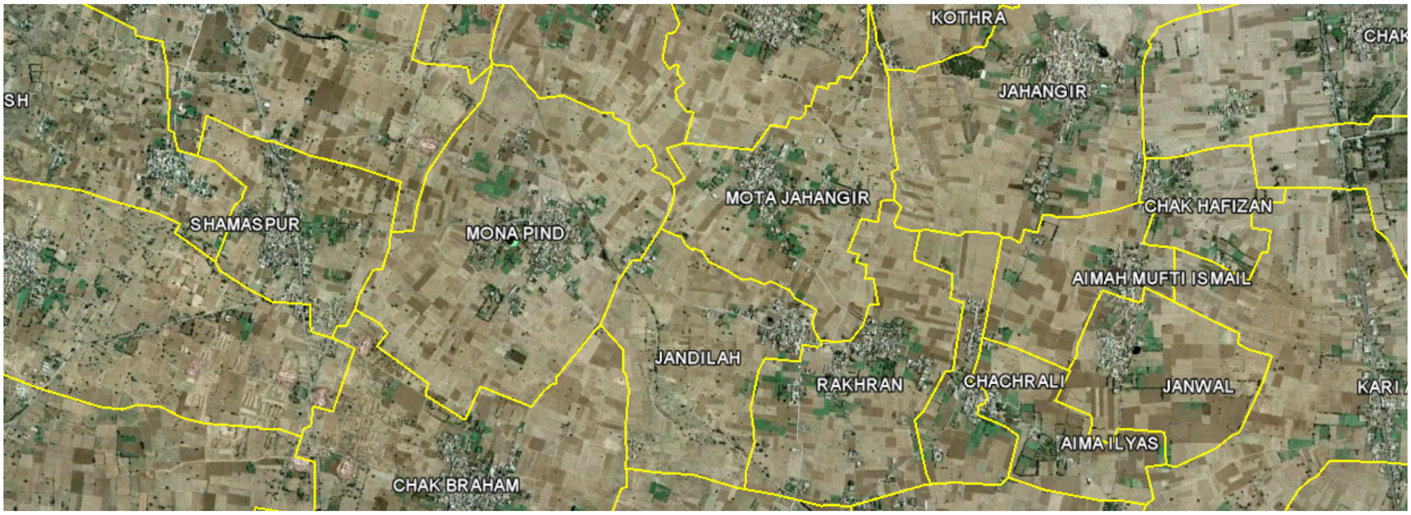
The Need: The significance of mauza-boundaries is immense. Most socio-economic and demographic data at the rural level, such as census data, national identity cards and patients reporting at a hospital, contain the mauza name. Yet these datasets currently are not geo-referenced, existing in tabular forms and not available on a map. Current maps of Survey of Pakistan show administrative boundaries only till Tehsil level. To understand the difference in coarseness, it may be mentioned that there are about 120 tehsils in Punjab but more than 26,000 mauzas. By building a mauza-level map of Tehsil Jhelum, we were able to show the census data on a geo-referenced map for the first time, enabling inference of spatial trends for more effective decision making.



[Top] Scan of a typical masavi, which is torn c a few places. [Bottom] Scan of a heavil damaged masavi.

Approach: Mauza boundaries are defined by the Land Revenue Department. The most recent revenue maps of District Jhelum were prepared in 1940 under the 3rd Settlement of the district. These maps, called *masavis*, are stored in the District Record Room and form the basis of all administrative boundaries at the rural level. For example, the jurisdiction of a *police thana*, *election halaqas*, *union councils*, *patwar circles* and *qanoongoees* are all defined through agglomeration of mauza boundaries. Yet, the maps are largely inaccessible, and some have decayed in the last 70 years to a point where they are unusable. While there could have been other approaches to extract the mauza boundaries, we identified the dire need of preservation of more than 5,000 masavis of District Jhelum as a design constraint, and started preparing a methodology for scanning these valuable documents.

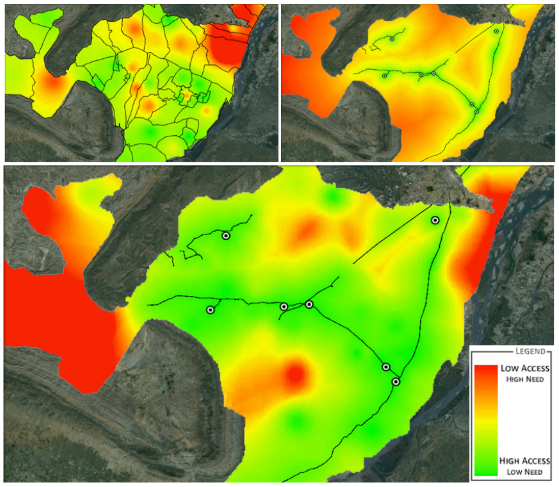
Methodology: Due to the age and the brittle nature of the paper on which the masavis are made, we did not scan them through a normal roller-scanner. Instead, we manufactured our own camera-based scanner which imaged each masavi in two halves. The scanning operation was completed within two weeks and generated approximately 10,000 images. Advanced image processing techniques were used to stitch the images together, first to combine the two halves of a masavi into a single image, then to mosaic the masavis of a whole mauza into one image and finally join images of adjacent mauzas together like a giant jigsaw puzzle. These steps required innovation in multiple directions: an understanding of the terminology of the patwari system and their mapping practices was generated, indigenous software was developed to speed up the mosaicing process and protocols for handling missing information, torn sheets and incomplete maps were created. Finally the mosaics of mauzas were geo-referenced on recent satellite imagery. We observed a high degree of alignment between the 70-year old masavis and today's satellite images. Seeing the similarity of features such as roads, fields and streams was a source of great satisfaction for our team; it not only confirmed the high accuracy of the survey procedures employed in the 1930s, but also provided verification that the methodology that we had employed to transfer these maps to digital form was indeed correct.



Portion of the mauza-boundary map overlay on a satellite image on Google Earth™

In addition to administrative boundaries, our team in partnership with the Population Census Organization also collected coordinates of infrastructure of interest, such as schools, religious places, government buildings and health facilities. This data was also integrated into the final product.

Potential Impact: The mauza-boundaries themselves are a very useful product for district administration, and our mauza maps were termed invaluable for the daily working of the district officials. However, their linkage with socio-economic data opens up the most exciting possibilities. As a demonstration, we developed a few applications of the use of mauza-boundaries for typical government functions. Census data was mapped to these boundaries using the numeric mauza-identifier available in both census and revenue records (*hadbast number*). To the best of our knowledge, this is the first time in Pakistan that it became possible to visualize census data on a map, hence enabling the inference of spatial trends. Furthermore, we also developed a spatial-analysis visualization to help the government plan the location of a new school or health unit. This was done by analysis of the mauza-level population, the existing location of these facilities and the accessibility of these facilities through roads [see inset]. An additional significant impact of the project is the availability of land revenue maps in a geo-referenced format, which will not only helps preserve the decaying original record, but also makes the land-revenue data much more readily accessible and useable.



Accessibility to Health Care: [Top Left] The spread of population, generated from mauza-level data projected from the 1998 census. [Top Right] Accessibility via road network to Basic Health Units. [Bottom] Combining these two factors allows us to highlight areas where a new BHU may be planned.

The synergistic combination of old 1940 maps and recent 2010 satellite imagery in this project has demonstrated a powerful combination for socio-economic planning in Pakistan. A detailed administrative boundaries map can indeed have many more potential uses, both for public and private sectors. It is, therefore, defined as one of the critical layers of any National Spatial Data Infrastructure (NSDI). Through this project, we have taken the first steps towards creating a comprehensive NSDI for Pakistan.

Case Studies

Mobile Phone Price as a Proxy for Socio-Economic Indicator

The frequency, granularity and accuracy of socio economic data is limited in many developing countries because of the cost, governance, and political economy challenges of data collection. For example, in Pakistan, it has not been possible to conduct the population census since the last 17 years. At the same time, the Information and Communication Technology (ICT) sector has grown enormously in Pakistan over the past decade, resulting in one of the highest teledensity levels in the region. This situation is not specific to Pakistan but is also prevalent in many developing countries.

The low cost, high granularity, real-time data trail left behind by mobile users of a region can be used as an alternate method to provide socio-economic insights and thus help policy makers plan and target resource allocations, evaluate ongoing interventions, and manage disasters.

For this study, we focus on a hitherto unexplored variable in mobile call records: handset models. From eleven months of data of prepaid subscribers in one district of Pakistan, we calculate average phone price of residents in the approximate region of each cellular tower, and then correlate it with variables in two publicly available socio economic datasets that provide village-level statistics.
We had access to eleven months (July 2013-June 2014) of cellphone usage data of the largest telecom provider in Pakistan, Mobilink, for District Jhelum. It was chosen because of its geographic and socio-economic heterogeneity. We used data for prepaid subscribers only, which make up for an overwhelmingly large percentage of total subscribers in Pakistan.

The data available for this study was not the actual Call Detail Records (CDRs). Instead, mobile usage variables for individual subscribers were aggregated on a daily basis. Separately, for each active day of each subscriber, we had a list of cell towers under which the subscriber performed any activity. This list of towers was not sorted in temporal order, thus did not allow us to track location during off-hours for identification of their residence. The dataset comprised of around 43 million records, a little over 360,000 unique users and 80 cell towers.

In our dataset, one of the variables was the handset model of the subscriber against which activity for the day was recorded. For these handset models, we manually collected their existing market prices, by searching popular retail websites in Pakistan. We managed to collect phone prices of around 85% of users. Handsets for which prices could not be found were almost all uncommon low-end handsets, for which the reported name did not match the name by which the model is marketed in the country. A fixed price of PKR 3000 (approximately USD 30) was used as their market value.

For models that had been discontinued and were not available in the marketplace, we collected their second-hand prices from local p2p online marketplaces. For users against which activity was recorded from more than one phone over the eleven months, we used the average price of the handset.

We used two publicly available socio economic census datasets in our study. Both datasets provide information aggregated for each Mauza. The first dataset is the Mauza Census of 2008, conducted by the Agricultural Census Organization (ACO). Our second dataset is the Population Census of 1998, conducted by the Population Census Organization (PCO). From PCO and ACO datasets, we used 21 and 213 socio-economic variables respectively.

To correlate mobile usage dataset against the socio-economic state of a region's residents, we need a method to approximate the place of residence of mobile users. We map users onto what we call their 'home towers', taken to be the tower under which their residence is located. Because of the unavailability of hourly data within a day, the most frequent locations of subscribers during off hours could not be tracked. Due to this limitation, we estimated the home tower to be the tower under which the subscriber was spotted the most during the eleven months.

In the mobile dataset, the towers were available to us as latitude-longitude pairs. To get an approximation of their area of coverage, we used Voronoi algorithm. The algorithm simply assigned region to each tower in such a manner that each point in the region is closer to its assigned tower than any other tower. Rather than directly mapping mauza statistics to cell towers, we first mapped census data onto settlements within a mauza boundary. This was done to improve the accuracy of the mapping of census variables to towers, because mauzas often cut across tower boundaries, while settlements, being sparse, are less likely to do so. Large settlements falling into multiple mauzas were split into smaller contiguous settlements.

As the next step, we took the census data already mapped onto settlements and aggregated it onto towers. We performed this mapping using weighted averages, based on the percentage of the total settled area in the tower belonging to the settlement.

Once census and mobile data variables were mapped onto towers, we computed two statistics as quantified measures of their relationships: Pearson's correlation coefficient R, capturing the general linear relationship between variables and *p*-value capturing probability of the null hypothesis being true. In our results, we consider two variables to be correlated only if $|R| > 0.35$ and $p < 0.01$.
Our results indicate that phone price correlates with a large number of socio-economic indicators. Though correlations with the individual variables may be moderate, the fact that cellphone price correlates so broadly and with indicators of varied socio economic theme goes to show that there is a clear relationship between the average phone price of a region and its socio economic state. Table below lists some socio economic variables (all correlated variables not listed), their theme and the Pearson correlation coefficient R and p-value with phone price.

Variable	**Theme**	**R Value**	**p-value**
Distance to College (Boys)	Education	-0.40	0.0026
Distance to College (Girls)	Education	-0.35	0.017
Distance to Child/Mother Center	Health	-0.40	0.0026
Distance to private MBBS doctor	Health	-0.43	$8.4\times10^{-0.5}$
Bricked streets	Infrastructure	-0.48	$8.6\times10^{-0.6}$
Bricked Drains	Infrastructure	-0.54	$4.4\times10^{-0.7}$
Construction Type of majority of houses	Infrastructure	-0.35	0.018
Toilet facilities	Hygiene	-0.62	9.4×10^{-10}
Distance to police station/post	Law & Order	-0.36	0.012
Distance to private veterinary facility	Livestock	-0.39	0.0044
Distance to govt. wheat/grain procurement center	Agriculture	-0.46	$2.6\times10^{-0.5}$
Distance to internet	Communications	-0.36	0.013
Availability of cable	Communications	0.38	0.0055

Distance to commercial bank	Financial	-0.48	$8.7\times10^{-0.6}$

Electricity Consumption

Objectives

The objective of the project is to spatially map out the electricity generation data across Pakistan. The pilot project was launched in Lahore and the plan is to scale up across the entire country. The project is not restricted to any particular purpose rather the idea is to get a spatial visualization of the electricity consumption data. Currently, the data available on electricity consumption is limited to excel sheets which provide detailed information regarding the consumption in each feeder. This project aims at mapping out this data in order to visually understand the distribution of electricity consumption across regions.

Benefits

The benefits of this project are multifold; as spatial mapping of electricity is important information for many organizations in the government and development sector. Some of the benefits for policy makers are as follows:

1. Valuable information for many proxy indicators such as poverty per capita as electricity consumption is a good indicator of the standard of living of households. Poverty distribution across the country can be traced through mapping which is otherwise ignored in statistics such as poverty per capita. This is because it fails to account for the variations across regions which can be captured through visual data.
2. Helps identify industrial areas and their electricity requirement.
3. Identifies underserved and neglected communities.
4. Can be used in conjunction with already existing maps and provide useful regional indicators.

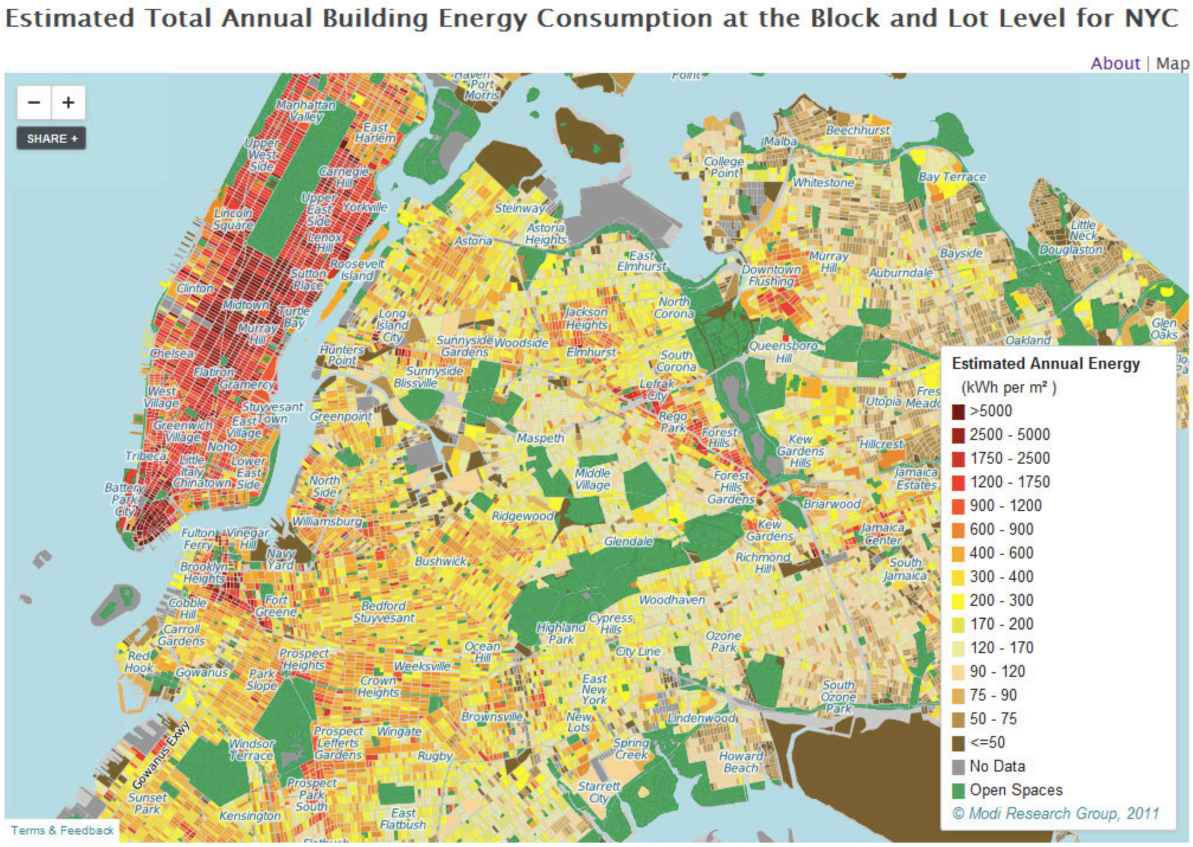
Currently the electricity department has access only to line diagrams of electricity consumption. Visual mapping can have multiple benefits for them as well. Some of them are as follows:

1. Better understanding of the electricity feeders installed.
2. Helpful in tracking electricity theft across regions. It will identify regions with a high theft rate and so on.
3. Identify consumption patterns of electricity regionally. Valuable information regarding electricity requirements across regions.
4. Identify electricity bill collection patterns regionally.
5. Identify temporal variations with regards to consumption. Particularly helpful in understanding change in electricity consumption patterns over time.

6. Provide information regarding load shedding; in particular identify areas which are affected the most and so on.

Examples

Examples of what the project aims to do are attached below.

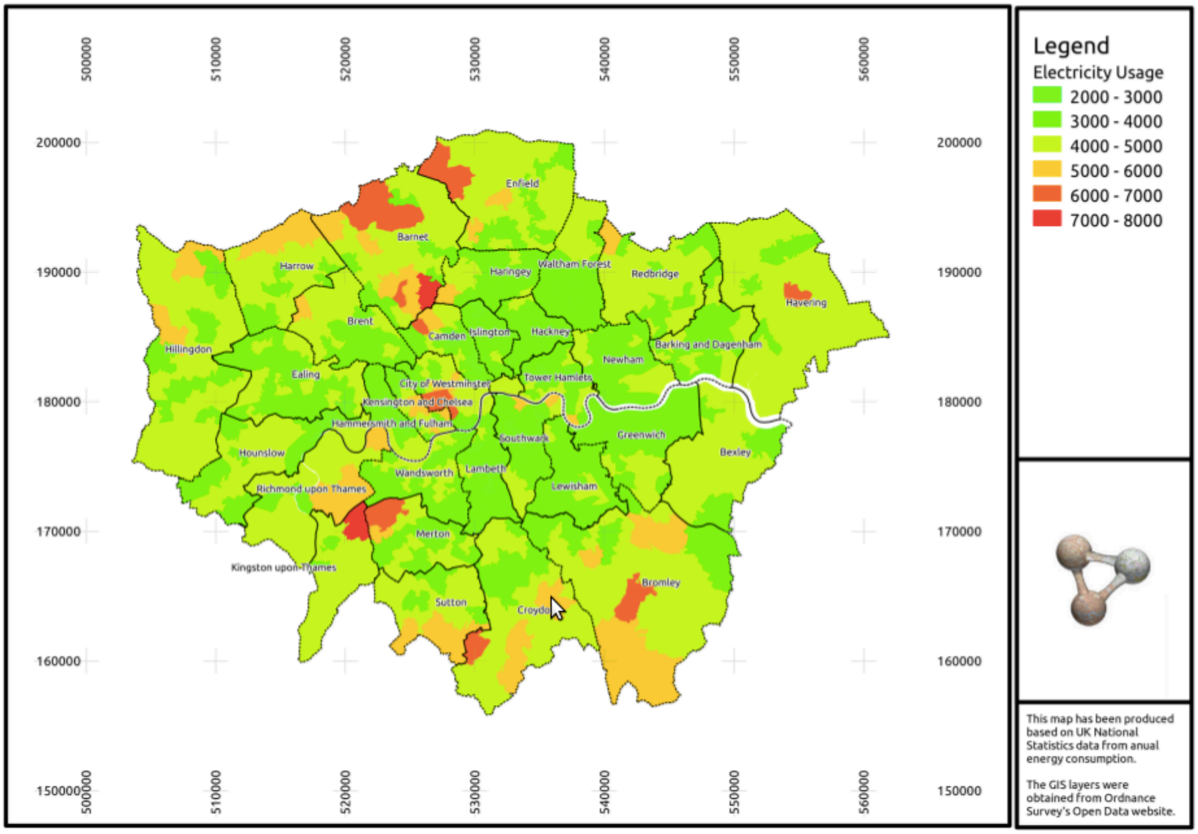


Estimated Total Annual Building Energy Consumption at the Block and Lot Level for NYC

Estimated Annual Energy
(kWh per m²)

🟫	>5000
🟥	2500 - 5000
🟥	1750 - 2500
🟥	1200 - 1750
🟧	900 - 1200
🟧	600 - 900
🟧	400 - 600
🟨	300 - 400
🟨	200 - 300
🟨	170 - 200
🟨	120 - 170
🟫	90 - 120
🟫	75 - 90
🟫	50 - 75
🟫	<=50
⬜	No Data
🟩	Open Spaces

© Modi Research Group, 2011



Problems and Approach

The major obstacles of this project are as follows:

1. There are too many addresses which have to be located manually.
2. Ambiguous and unavailable addresses on the map
3. Addresses do not follow any standard procedure.

Given these obstacles and the fact that the project has to be scaled up to entire country, there was a need to develop an automated system which could process these addresses. Since the addresses cannot be located on the map because of the poor records, the approach has switched from identifying addresses to identifying regions. That is, maps would show the variations in electricity consumption across regions rather than variations across households.

Hierarchy of regions

The city of Lahore can be divided at 3 levels:-

1. Circles: can be divided into 5 circles
2. Substations
3. Feeders: approximately 1200 feeders

4. Approximately 3 million consumers at the district level.

For the purpose of our project, the unit of analysis is the electricity feeder.

Challenges in Automation

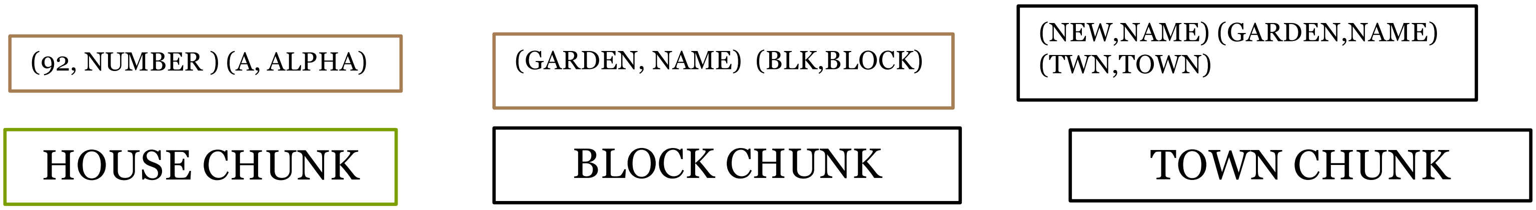
Major problems with the raw addresses were as follows:

1. Spelling variations. For example, Jeevan, Jiwan, Jewan, Geven, Jevn are all variations of the same word.
2. Incomplete addresses. For example, Jiwan Hana, LHR. Here, no address is available, just the block name.
3. Inconsistent Abbreviations. For example, New G/Town, G-Town, N G Town, N G are all variations of the New Garden Town.
4. Data Entry Errors. For example, (, BLKLHR, LK, BLKLH, HANALHR. Missing a single space or usage of the incorrect bracket sign causes problems in the automation of the system.

Automated system

The automated system can be divided into 4 steps:

1. The raw addresses are processed in order to identify the high frequency words within the feeder. For example, "BL" and "TWN" are words that are found often in addresses.
2. Metaphone is applied on these high frequency words in order to correct their spellings. For example, the word "TWN" is replaced by TOWN or GARDAN is replaced by GARDEN. The metaphone is an algorithm which identifies the high frequency words, creates a dictionary by grouping together all the words with slight spelling variations and standardizes the spellings by replacing them with the highest frequency word. The highest frequency word is assumed to be the right spelling.
3. Assigning each word in the address with a tag of its type. For example, "92 - A GARDAN BL NEW GARDEN TWN LHR" would look like this after the two steps: "(92, NUMBER) (A, ALPHA) (GARDEN, NAME) (BL,BLOCK) (NEW,NAME)(GARDEN,NAME) (TOWN,TOWN) (LHR,CITY)"
4. Finally the addresses are extracted in chunks automatically. For example:


(92, NUMBER) (A, ALPHA)

(GARDEN, NAME) (BLK,BLOCK)

(NEW,NAME) (GARDEN,NAME) (TWN,TOWN)

HOUSE CHUNK

BLOCK CHUNK

TOWN CHUNK

Problems from the initial mapping

1. Some feeders overlap in the service area covered by them. That is, one part of an area could be served by say 2 feeders.
2. Some feeders also serve discontinuous areas. For example, some areas are non-contiguous yet covered by the same feeder.

Conclusion and Future Work

This paper presents a study trying to understand the relationship between cellular usage data of a region and its socio economic state. A particular finding of interest is that average phone price of a region is highly correlated with its socio economic indicators. We also show that variables related to expenditure also exhibit significant correlations. Thus we make the case that cellular usage data is a viable alternative or proxy for socio economic standing of a region. In addition, its real time availability and low cost make it ideal for policy makers in planning resource allocations, evaluating ongoing interventions and for disaster management. This study shows that the average mobile phone price is a good proxy indicator for several socio-economic indicators of a region.

Future work would focus on trying to correlate with other socio economic datasets particularly poverty data collected by a national poverty support program (Benazir Income Support Program) which is not only much more recent but is also granular down to the level of individual households and focuses on economic variables including income. For future work, we also want to include other carriers to remove any bias any particular carrier may introduce in data.

Annex I

Web Based Interactive Tool

All these data-sets, which were available in raw numbers in excel sheets were mapped out using the online interactive tool. The details of this project are outlined below.

Objectives and Introduction

The objective of this project is to essentially map out important socio-economic indicators across regions in Pakistan. These indicators are available in the form of cold, raw figures in excel files with limited usage for policy makers and users. The idea behind the spatial mapping is to have a more improved and nuanced understanding of the different indicators regionally. The benefits of this are multifold; particularly in increasing the accessibility of such indicators across masses and amongst policy makers. This will of course raise the general public awareness regarding the distribution of facilities and also aid policy makers in making better informed decisions.

In doing so, the tool made used of several data sets such as the Punjab Federal Statistics, Multiple Indicator Cluster Survey, Pakistan Social and Living Standards Measurement and so forth. The data in these data sets can divided into 3 levels of granularity: District, Tehsil and Mauza. Our aim was to map out the data at the lowest level of granularity possible in order to get a more precise picture regarding the regional variations in Pakistan.

This report will provide a run through to the interactive tool by looking into the various indicators being covered. It will then go on to highlight the difficulties and shortcomings of the online interactive tool. The report will end with a look towards the future; that is, how we envision the tool is utilized by people.

Interface of the Tool

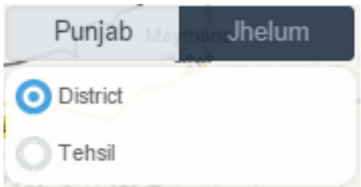
This section will look into the interface of the tool and highlight its key features.

The tool allows visualizations at four levels of granularity.

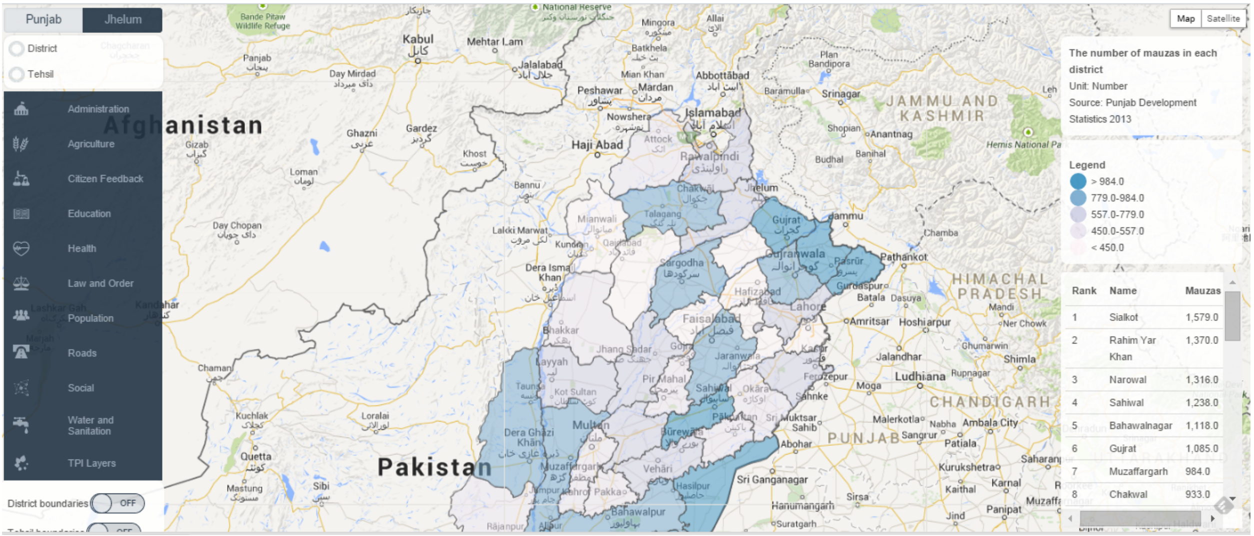
District

Tehsil

Mauza



TPI Poverty Mapping



The number of mauzas in each district

Unit: Number

Source: Punjab Development Statistics 2013

Legend

● > 984.0

● 779.0-984.0

● 557.0-779.0

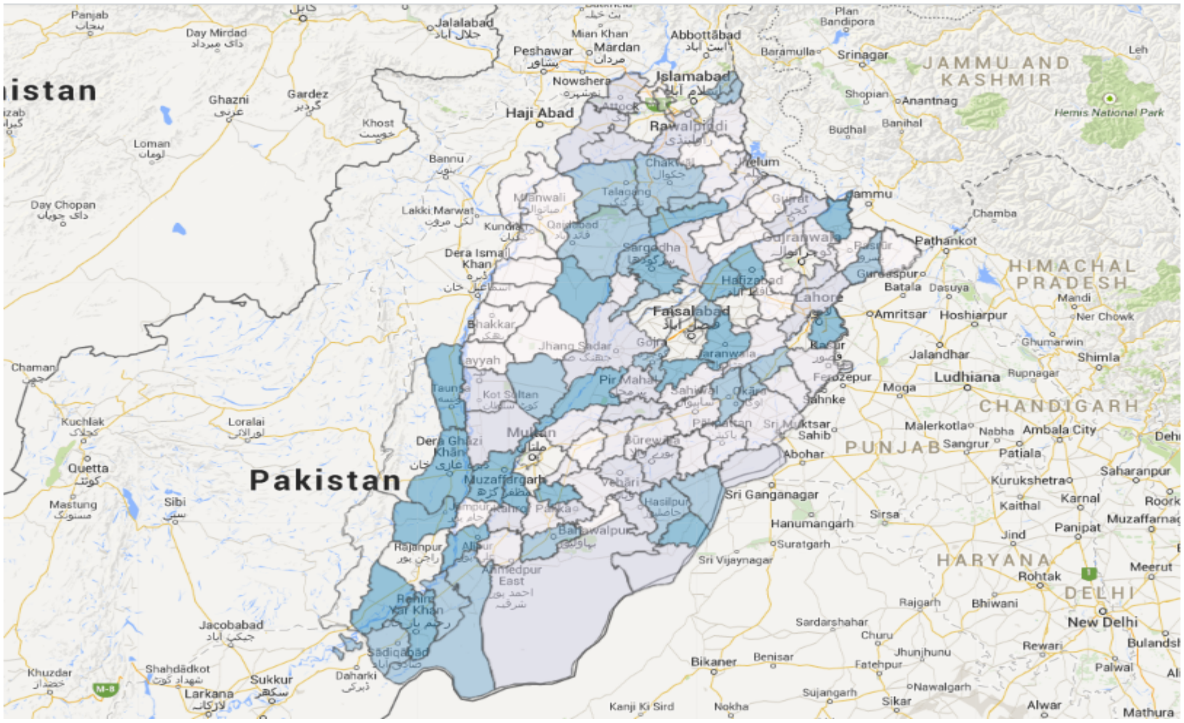
● 450.0-557.0

● < 450.0

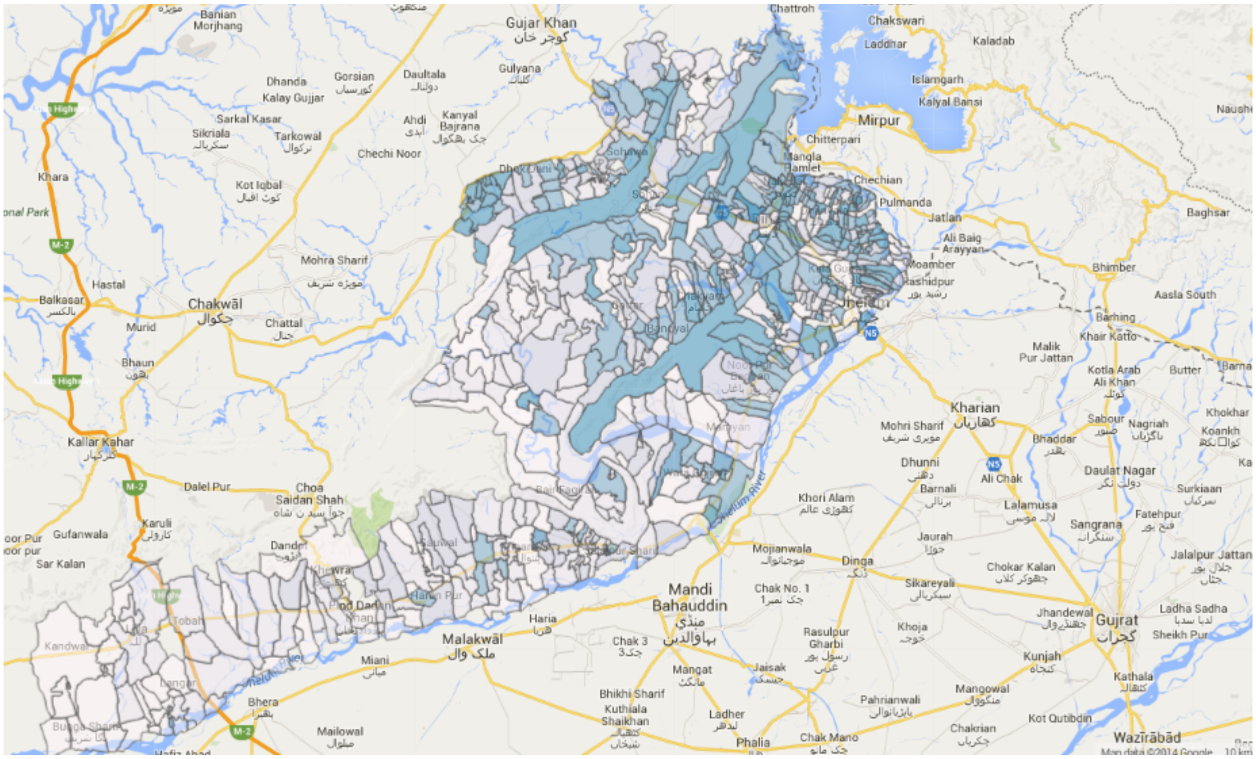
Rank	Name	Mauzas
1	Sialkot	1,579.0
2	Rahim Yar Khan	1,370.0
3	Narowal	1,316.0
4	Sahiwal	1,238.0
5	Bahawalnagar	1,118.0
6	Gujrat	1,085.0
7	Muzaffargarh	984.0
8	Chakwal	933.0

The image above shows the complete view of the tool with view set at Punjab and granularity at Districts.

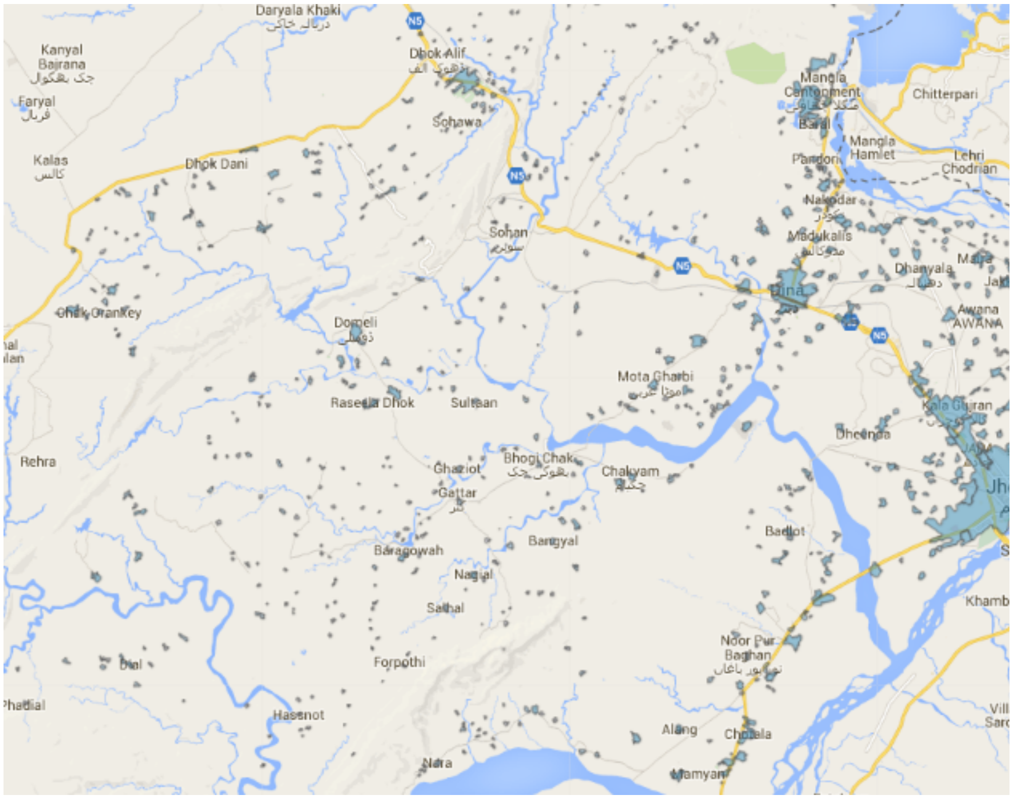
View and granularity can be selected from menu at the top left. Punjab View allows data to be viewed at granularity District or Tehsil. Jhelum view on the other hand allows visualization at Mauza and Settlements granularity. Following screenshots contain all views and granularities.

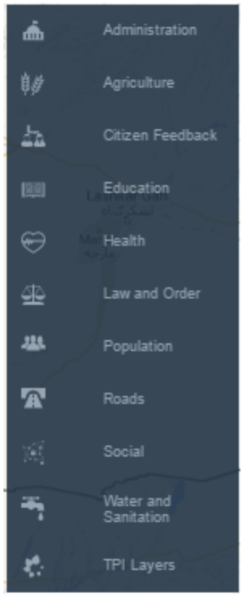


Pakistan
istan
Jalalabad
Peshawar
Mardan
Abbottābad
Srinagar
JAMMU AND KASHMIR
Islamabad
Nowshera
Haji Abad
Rawalpindi
Hemis National Park
Leh
Bannu
Chakwal
Jhelum
Jammu
Lakki Marwat
Mianwali
Talagang
Gujrat
Chamba
Pathankot
HIMACHAL PRADESH
Dera Ismail Khan
Sargodha
Gujranwala
Gurdaspur
Batala
Mandi
Faisalabad
Hafizabad
Lahore
Amritsar
Hoshiarpur
Bhakkar
Jhang Sadar
Kasur
Jalandhar
Shimla
Layyah
Ferozepur
Ludhiana
Chaman
Kot Sultan
Pir Mahal
Sahiwal
Okara
Moga
CHANDIGARH
Kuchlak
Loralai
Dera Ghazi Khan
Multan
Pakpattan
Sri Muktsar Sahib
Malerkotla
Ambala City
Quetta
Burewala
Abohar
PUNJAB
Patiala
Muzaffargarh
Vehari
Kurukshetra
Saharanpur
Mastung
Sibi
Hasilpur
Sri Ganganagar
Kaithal
Karnal
Bahawalpur
Hanumangarh
Sirsa
Jind
Panipat
Rajanpur
Suratgarh
HARYANA
Rohtak
Meerut
Ahmedpur East
Sri Vijaynagar
Rahim Yar Khan
DELHI
New Delhi
Jacobabad
Sardarshahar
Rajgarh
Bhiwani
Sadiqabad
Churu
Rewari
Khuzdar
Shahdādkot
Sukkur
Daharki
Bikaner
Benisar
Jhunjhunu
Fatehpur
Palwal
Larkana
Sujangarh
Nawalgarh
Sikar
Alwar
Kanji Ki Sird
Nokha
Mathura



Banian Morjhang
Gujar Khan
Chattroh
Chakswari
Kaladab
Laddhar
Dhanda Kalay Gujjar
Gorsian
Daultala
Gulyana
Islamgarh
Kalyal Bansi
Naushera
Sarkal Kasar
Sikriala
Tarkowal
Ahdi
Kanyal Bajrana
Mirpur
Chechi Noor
Chitterpari
Khara
Kot Iqbal
Mangla Hamlet
Chechian
Pulmanda
Jatlan
Baghsar
Ali Baig Arayyan
Mohra Sharif
Moamber
Bhimber
Hastal
Rashidpur
Aasla South
Balkasar
Chakwāl
Chattal
Murid
Barhing
Khair Katto
Malik Pur Jattan
Bhaun
Kotla Arab Ali Khan
Butter
Kharian
Kallar Kahar
Mohri Sharif
Sabour
Nagriah
Khokhar
Koankh
Bhaddar
Dhunni
Dalel Pur
Choa Saidan Shah
Ali Chak
Daulat Nagar
Barnali
Khori Alam
Surkiaan
Karuli
Lalamusa
Fatehpur
Gufanwala
Sangrana
Dandot
Mojianwala
Dinga
Jaurah
Sar Kalan
Jalalpur Jattan
Mandi Bahauddin
Chak No. 1
Chokar Kalan
Sikareyali
Pind Dadan Khan
Haria
Ladha Sadha
Rasulpur Gharbi
Khoja
Jhandewal
Gujrat
Sheikh Pur
Malakwāl
Chak 3
Kandwal
Miani
Kunjah
Jaisak
Mangat
Mangowal
Kathala
Bhera
Bhikhi Sharif
Kuthiala Shaikhan
Ladher
Pahrianwali
Chakrian
Kot Qutibdin
Mailowal
Phalia
Chak Mano
Wazīrābād
Hafiz Abad
Jhelum River
M-2
N5
Map data ©2014 Google 10 km

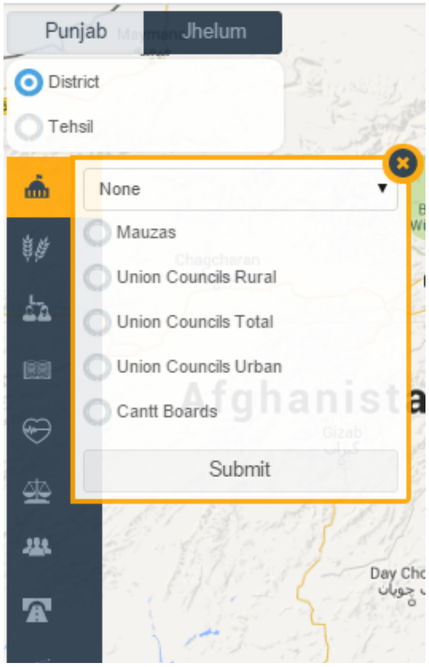




Each view and granularity pair has its own variable categories under which we've nested different variables. These categories correspond to different socio economic themes such as Health, Education, Law order etc.

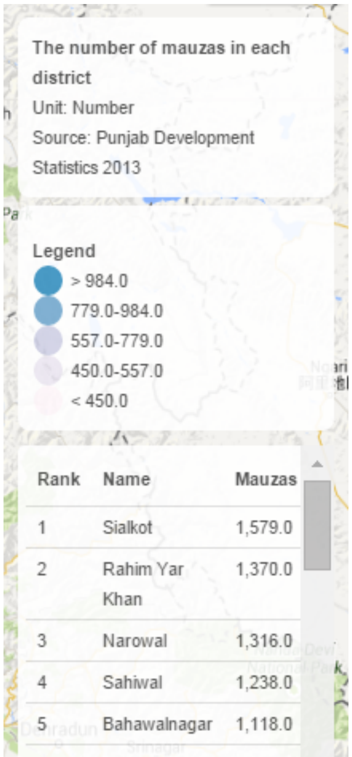
To the right is view of the variable categories for Districts. Notice the added icons for each category to increase comprehensibility.

On selecting a category, the menu collapses, now with just the icons visible, with icon of the selected category highlighted and a submenu slides into view with the nested variables.



The submenu lists all the variables corresponding to the selected category. In order to view a variable on the map, the user only needs to select its associated radio button and press submit button.

In order to go back to the old view the user can either press the cross button or select any icon in the collapsed menu.



The number of mauzas in each district

Unit: Number

Source: Punjab Development Statistics 2013

Legend

● > 984.0

● 779.0-984.0

● 557.0-779.0

● 450.0-557.0

● < 450.0

Rank	Name	Mauzas
1	Sialkot	1,579.0
2	Rahim Yar Khan	1,370.0
3	Narowal	1,316.0
4	Sahiwal	1,238.0
5	Bahawalnagar	1,118.0

For each variable, to the right are a set of menus providing important information regarding the selected variable.

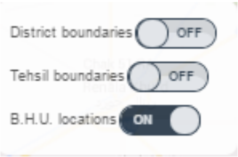
At the top is an info box, providing vital information about what's being shown on the map. This information includes complete name of the variable, the unit of the value of its variables and most importantly, its source census or survey along with the year at which the census was conducted.

Below the info box is the **Legend**. The Legend explains how the value of the variables were binned and what color was assigned to each bin. Darker shades have been used for bins capturing the higher end of the spectrum and vice verse.

At the bottom to the right is a **Rankings Table**. This table has three columns: Rank, Name of the units and the value of the variable for each unit. This table is sorted in descending order.

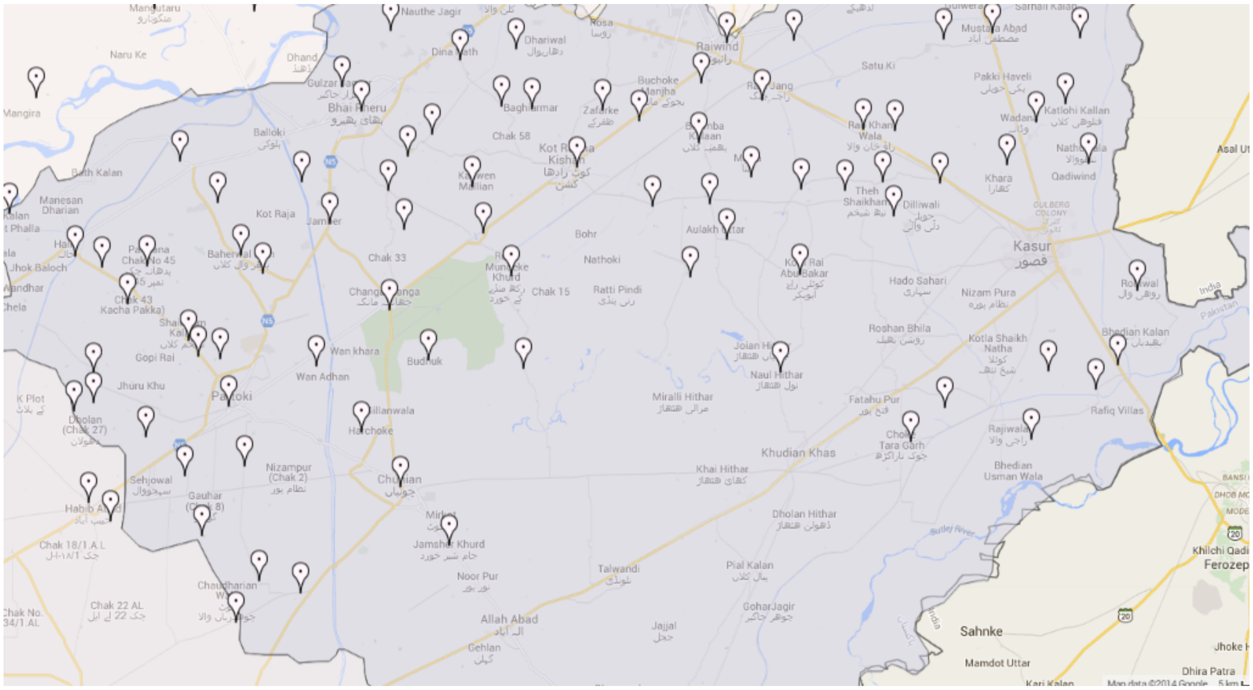
Toggle Layers

Another important feature of the app are toggle layers that can toggled on or off at any time, regardless of the view or granularity selected. Currently the layers we have on the tool are District boundaries, tehsil boundaries and BHU locations.



District boundaries OFF
Tehsil boundaries OFF
B.H.U. locations ON

The image below shows BHU Locations turned on with District granularity.



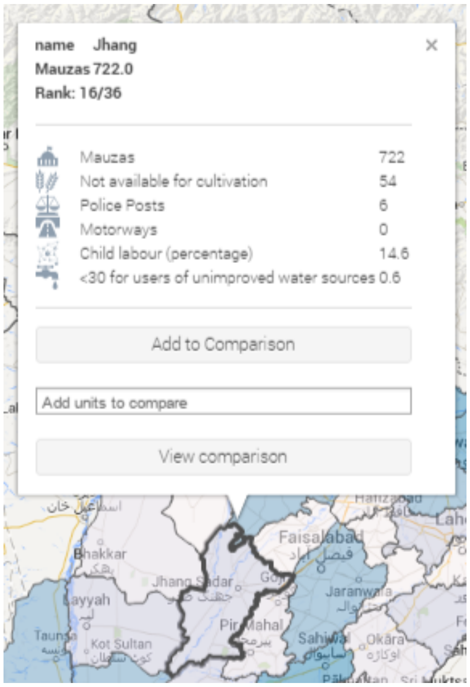
Popup window

When a unit is selected, the boundary of that unit is highlighted and a popup window appears providing details of the unit selected.

Details on the info window include

- name of the unit
- value of the variable selected
- how it ranks in comparison to other units
- values of one important variable from each category for the selected granularity

At the bottom are UI components that can be used to compared units



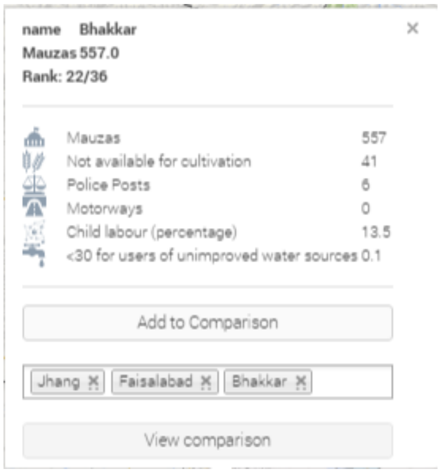
name **Jhang**
Mauzas 722.0
Rank: 16/36

Mauzas	722
Not available for cultivation	54
Police Posts	6
Motorways	0
Child labour (percentage)	14.6
<30 for users of unimproved water sources	0.6

Comparison

The tool also allows the user to compare as many units as he wants.

Comparisons take place via info window. When the user selects a unit, he can press the 'Add to Comparison button' to add the selected unit to compare. Similarly, the user can open info window of other units that are to be compared. When all the units have been selected, pressing the View Comparison button can take to the Advanced view containing graphs comparing the selected units.



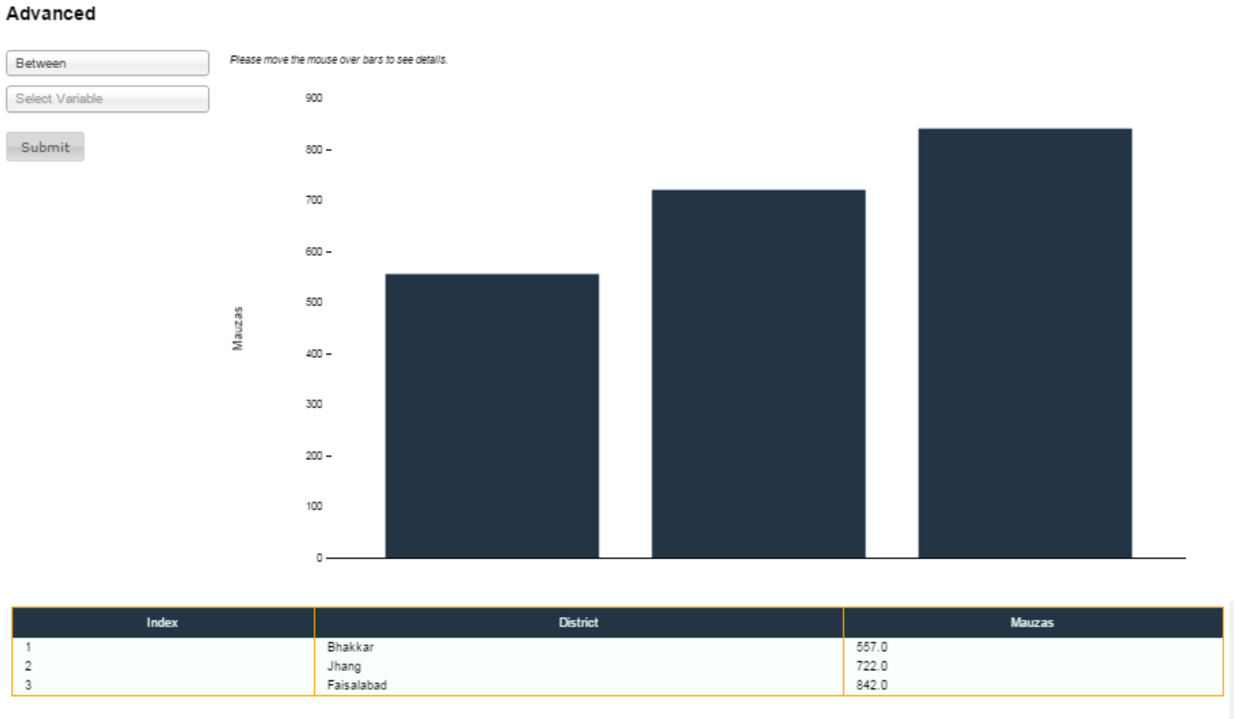
name **Bhakkar**
Mauzas 557.0
Rank: 22/36

Mauzas	557
Not available for cultivation	41
Police Posts	6
Motorways	0
Child labour (percentage)	13.5
<30 for users of unimproved water sources	0.1

The user can also deselect units by pressing the small cross button at the top right of the name of the unit.

Pressing the View Comparison button results in opening of the advanced comparison view. This view primarily consists of bar graphs of selected units with a table below sorted by variable value in descending order.

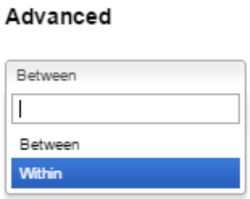
Moving the cursor over a bar displays name of the unit.

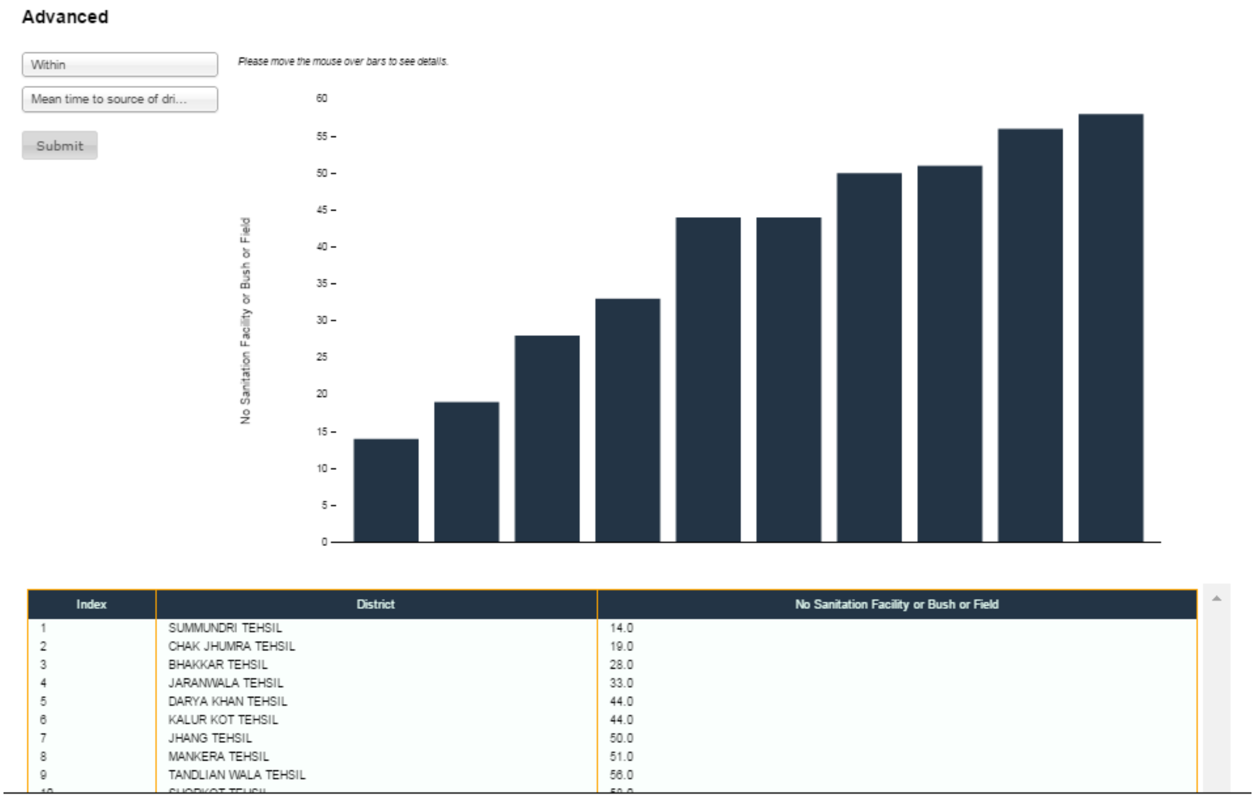


Please move the mouse over bars to see details.

Index	District	Mauzas
1	Bhakkar	557.0
2	Jhang	722.0
3	Faisalabad	842.0

Another advanced feature we have in the tool is allowing comparisons within units. So for instance if the user was comparing districts initially and from the dropdown selects within. The bar graph would update itself to move display comparison at a lower granularity, for the originally selected units.

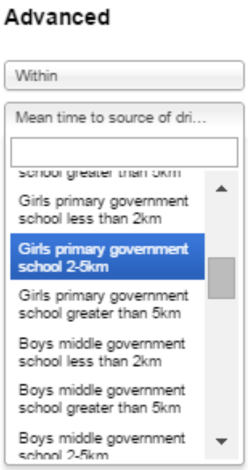



Advanced
Within
Mean time to source of dri...
Submit
Please move the mouse over bars to see details.
No Sanitation Facility or Bush or Field
60
55
50
45
40
35
30
25
20
15
10
5
0

Index	District	No Sanitation Facility or Bush or Field
1	SUMMUNDRI TEHSIL	14.0
2	CHAK JHUMRA TEHSIL	19.0
3	BHAKKAR TEHSIL	28.0
4	JARANWALA TEHSIL	33.0
5	DARYA KHAN TEHSIL	44.0
6	KALUR KOT TEHSIL	44.0
7	JHANG TEHSIL	50.0
8	MANKERA TEHSIL	51.0
9	TANDLIAN WALA TEHSIL	56.0
10	SHORKOT TEHSIL	58.0

We also allow the user to change variables in the advanced view.

So if the user opened the advanced view with a variable X selected on the map. In order to compare the same units, but on a different variable, the user does not have to go back and start the process all over again. All that is required is to just select another variable from the variables dropdown to the right.


Advanced
Within
Mean time to source of dri...
school greater than 5km
Girls primary government school less than 2km
Girls primary government school 2-5km
Girls primary government school greater than 5km
Boys middle government school less than 2km
Boys middle government school greater than 5km
Boys middle government school 2-5km

Problems

- The biggest problem in this project was regarding the lack of data at the granular level. For example, the MICS of 2010 did not have data available at the Tehsil level. Therefore, we had to make use of the data from 2006 which was available at the Tehsil level.
- Even when data was available, it was in a form which could not be used for processing. In other words, it had to digitized in the form of excel sheets before it could be of use. This was an extremely time consuming exercise of putting all the data into excel sheets.
- Mauza boundaries themselves were not easily available. Even in the rare case that they were, there were contradictions between different sources

and sometimes overlapping mauzas as well. Lack of demarcated boundaries of the mauzas was of course an impediment.

Because of the reasons stated above, we choose Jhelum as a model district because data at such fine granularity was not available for other districts.

Future

The initial aim of the project was that this data could be utilized by policy makers who would then provide us with periodic feedback regarding the system. Through this feedback, the system could be modified and expanded in ways which would be useful to policy makers. We hope to get feedback from policy makers regarding the utility of the project which would facilitate future improvements.