Working paper

International Growth Centre

Mapping the grid

Sohaib Ahmad Khan
Ijlal Hussain Naqvi

April 2016

DIRECTED BY
LSE
UNIVERSITY OF OXFORD

FUNDED BY
UKaid
from the British people

Mapping the Grid

Table of Contents

Executive Summary

The electricity distribution system is inherently spatial in nature but a spatial representation of the system does not exist. Such a representation can help power managers develop a spatial understanding of the system helping them to make better strategic and resource allocation decisions.

Apart from aiding in management of the distribution system the data can be used to infer socio economic status of an area based on its power consumption. It has been established power consumption has a bi-directional causality with socio-economic status. Currently socio-economic indicators may be available at the tehsil level, however, this project's output is much more granular. In addition, power consumption information can be compiled quickly compared to census information providing a method to infer the economic changes in areas.

Power sector officials have started to recognize the importance of such a representation leading some distribution companies to start geo-tagging of meter reading activities.

Currently power managers can see power consumption information as a list, however, it is impossible to make spatial inferences from a list. This project has taken a different approach – one that does not rely on collecting information from the ground but instead uses existing data sources to create the spatial representation. This results in certain approximations.

The spatial representation developed as part of this project is not complete. This is because it relies on existing geo-referenced databases, which are incomplete for Pakistan. This is especially true for areas on the outskirts of the city or areas that are not zoned properly or known informally.

However, what we have been able to show even in the absence of a complete geo-referenced database we can map a large percentage of the feeders using this innovative approach. Of 259 feeders in Circle V of Lahore we were able to map 102 completely.

The Feeder

A feeder is the physical wire that transmits electricity from grid stations to consumers. It is also the administrative unit used by the distribution companies who view this as the base unit for their analysis. Many indicators, such as line losses, are calculated and viewed at the feeder level for management purposes. Therefore it is useful and appropriate to map the grid at the level of a feeder. It is of course theoretically possible to have a spatial representation of the grid at the household level, however this is not practical and may not be very useful.

Distribution of Feeders in Lahore

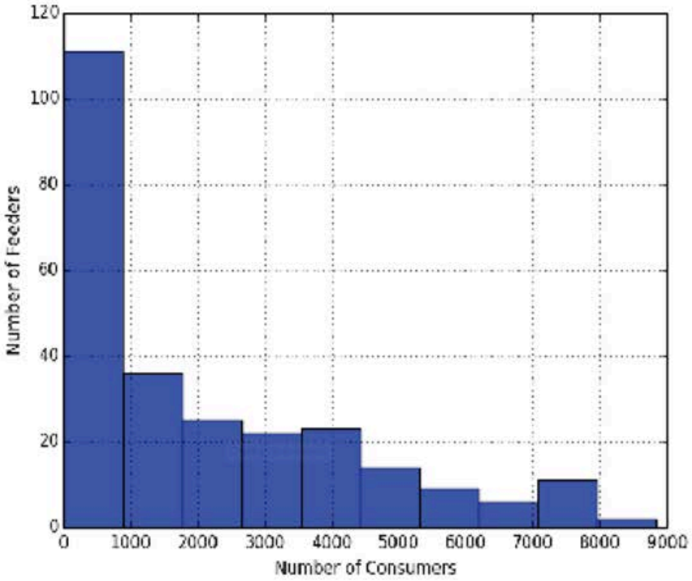


Fig 1. The figure above shows the distribution of feeders by the number of consumers served.. Most feeders serve under 2000 consumers. There are a few feeders serving a large number of consumers and generally serve areas considered to have a high population density

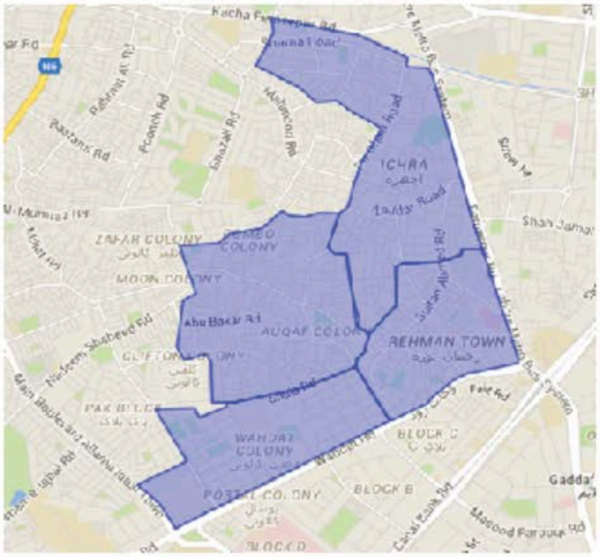


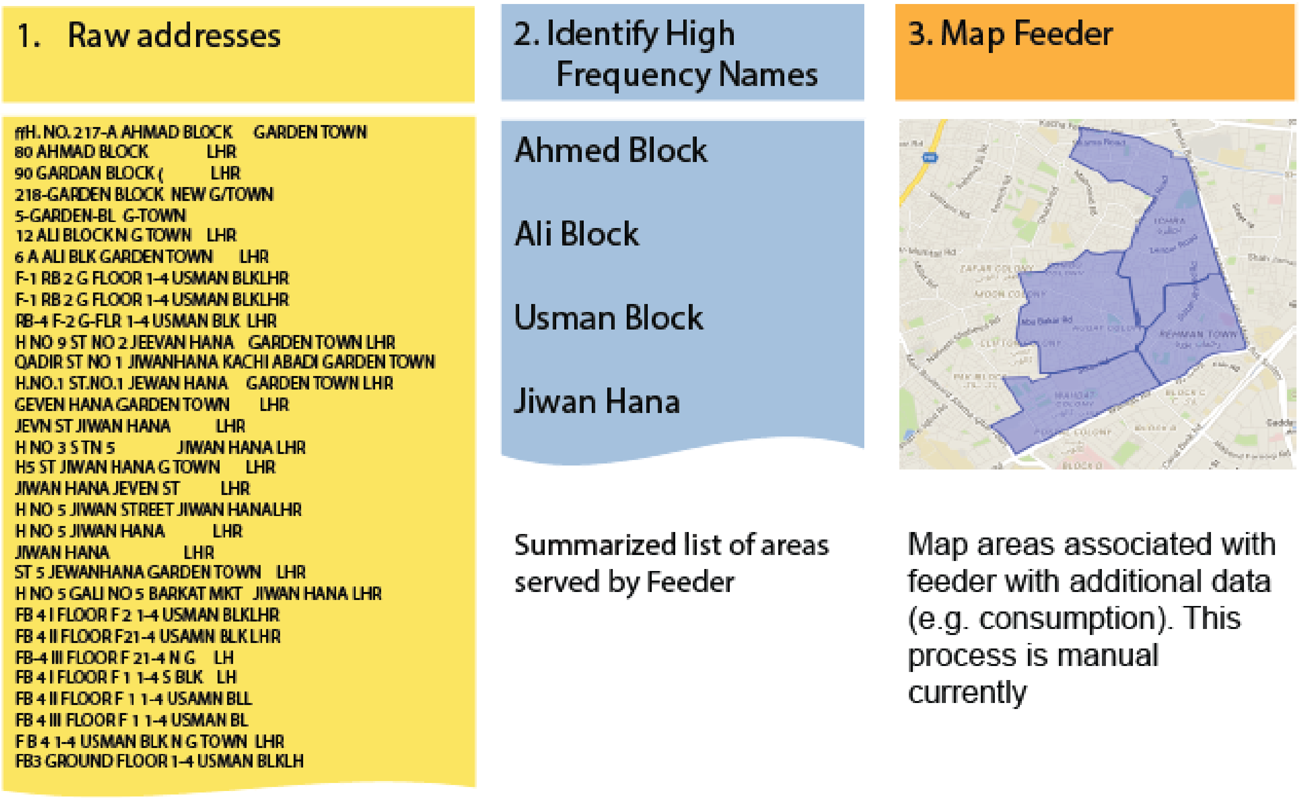
Fig 2. The footprint of a feeder 103. The four areas that this feeder serves are mapped separately. These are (counter clockwise from t top right) Icchra, Rehman Pura, Wahdat Colony and Shah Kamal. Like this feeder most feeders serve contiguos areas

The Mapping Process

We were provided a list of consumer addresses against each feeder for Circle V of Lahore. From names of areas served by each feeder were extracted and then geo-referenced. Taken together the footprints of these areas would then form the footprint of a feeder.

One option was to go through these addresses manually and extract the areas served by each feeder. Although possible this approach would take a significant amount of manual effort. An alternative approach was to develop an algorithm that would automatically extract key areas for a feeder given a list of addresses for that feeder. Since we want to scale this process to other circles of Lahore and possibly other cities we decided to automate the process as much as possible. An algorithm was developed that takes as input a list of addresses for each feeder and outputs the names of key areas served by these feeders. These areas are then geo-coded manually.

An Overall View of the Mapping Process



1. Raw addresses

ffH. NO. 217-A AHMAD BLOCK GARDEN TOWN
80 AHMAD BLOCK LHR
90 GARDAN BLOCK (LHR
218-GARDEN BLOCK NEW G/TOWN
5-GARDEN-BL G-TOWN
12 ALI BLOCK N G TOWN LHR
6 A ALI BLK GARDEN TOWN LHR
F-1 RB 2 G FLOOR 1-4 USMAN BLKLHR
F-1 RB 2 G FLOOR 1-4 USMAN BLKLHR
RB-4 F-2 G-FLR 1-4 USMAN BLK LHR
H NO 9 ST NO 2 JEEVAN HANA GARDEN TOWN LHR
QADIR ST NO 1 JIWANHANA KACHI ABADI GARDEN TOWN
H.NO.1 ST.NO.1 JEWAN HANA GARDEN TOWN LHR
GEVEN HANA GARDEN TOWN LHR
JEVN ST JIWAN HANA LHR
H NO 3 S TN 5 JIWAN HANA LHR
H5 ST JIWAN HANA G TOWN LHR
JIWAN HANA JEVEN ST LHR
H NO 5 JIWAN STREET JIWAN HANALHR
H NO 5 JIWAN HANA LHR
JIWAN HANA LHR
ST 5 JEWANHANA GARDEN TOWN LHR
H NO 5 GALI NO 5 BARKAT MKT JIWAN HANA LHR
FB 4 I FLOOR F 2 1-4 USMAN BLKLHR
FB 4 II FLOOR F21-4 USAMN BLK LHR
FB-4 III FLOOR F 21-4 N G LH
FB 4 I FLOOR F 1 1-4 S BLK LH
FB 4 II FLOOR F 1 1-4 USAMN BLL
FB 4 III FLOOR F 1 1-4 USMAN BL
F B 4 1-4 USMAN BLK N G TOWN LHR
FB3 GROUND FLOOR 1-4 USMAN BLKLH

2. Identify High Frequency Names

Ahmed Block

Ali Block

Usman Block

Jiwan Hana

Summarized list of areas served by Feeder

3. Map Feeder

Map areas associated with feeder with additional data (e.g. consumption). This process is manual currently

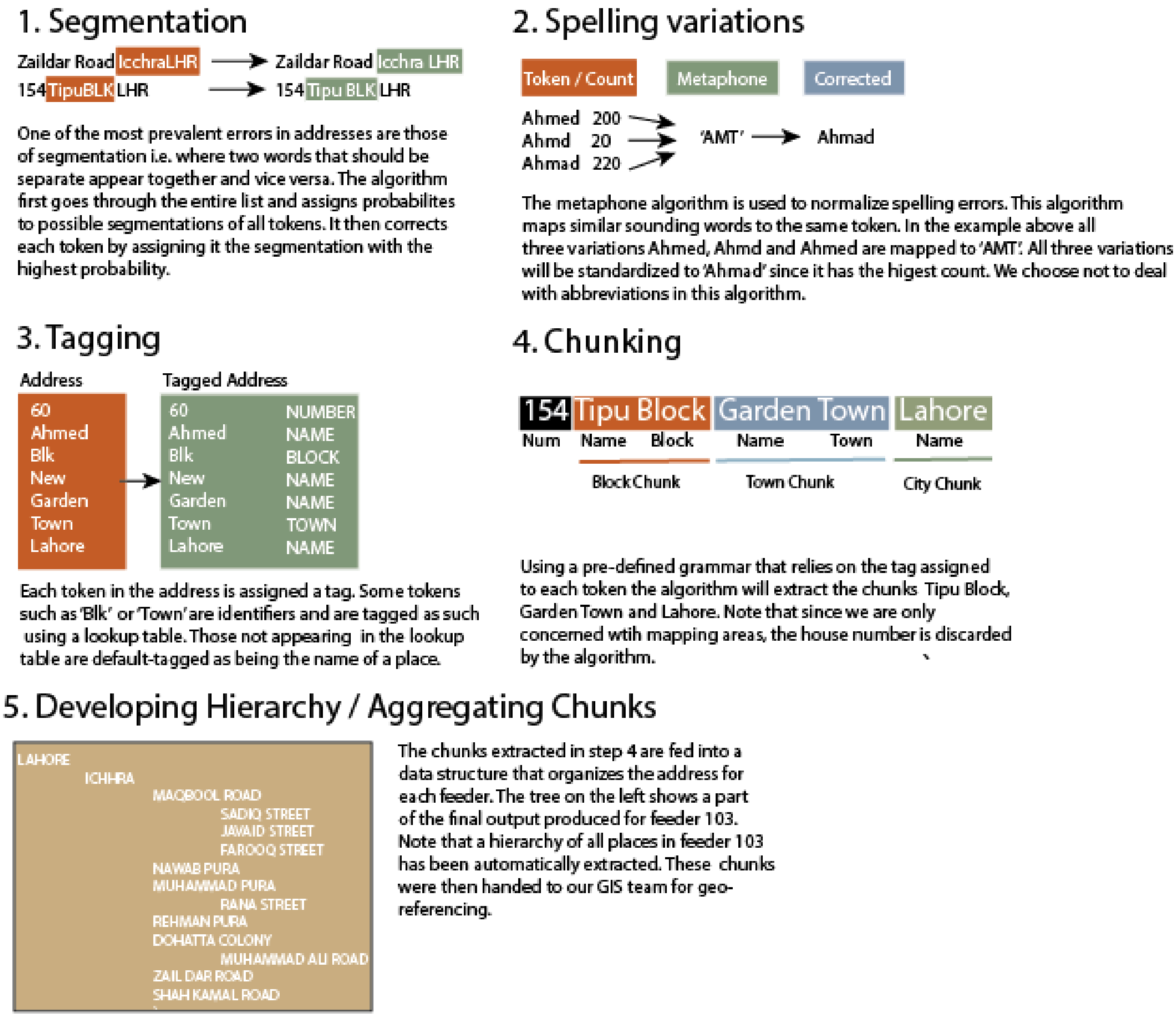
The figure above shows an overall view of the process used to obtain the footprint of a feeder. In step 1 a list of raw addressed for a feeder is fed into the system. In step 2 the system extracts key areas from this list of addresses, which are then mapped to obtain the footprint of the feeder.

From Addresses to Chunks

This step of the process takes as input raw addresses for a feeder and outputs a list of key areas associated for that feeder. These will be referred to as chunks. The number of chunks extracted for each feeder by the algorithm can vary from around 5 to about 20 depending upon the number of different areas that it serves.

Chunks are extracted from each address. As the algorithm goes through the address list it maintains counts of all the different chunks encountered. Chunks having counts over a threshold are identified as key areas for that feeder. The main challenge to automation is the amount of variations that are apparent in the addresses. These variations are those of structure, spellings, segmentation and inconsistent use of abbreviations.

Address Processing Algorithm



One of the most prevalent errors in addresses are those of segmentation i.e. where two words that should be separate appear together and vice versa. The algorithm first goes through the entire list and assigns probabilites to possible segmentations of all tokens. It then corrects each token by assigning it the segmentation with the highest probability.

The metaphone algorithm is used to normalize spelling errors. This algorithm maps similar sounding words to the same token. In the example above all three variations Ahmed, Ahmd and Ahmed are mapped to 'AMT'. All three variations will be standardized to 'Ahmad' since it has the higest count. We choose not to deal with abbreviations in this algorithm.

Each token in the address is assigned a tag. Some tokens such as 'Blk' or 'Town' are identifiers and are tagged as such using a lookup table. Those not appearing in the lookup table are default-tagged as being the name of a place.

Using a pre-defined grammar that relies on the tag assigned to each token the algorithm will extract the chunks Tipu Block, Garden Town and Lahore. Note that since we are only concerned wtih mapping areas, the house number is discarded by the algorithm.

The chunks extracted in step 4 are fed into a data structure that organizes the address for each feeder. The tree on the left shows a part of the final output produced for feeder 103. Note that a hierarchy of all places in feeder 103 has been automatically extracted. These chunks were then handed to our GIS team for geo-referencing.

This algorithm first divides the address string into tokens, which are individual words in an address. These are then run through to correct for variations due to segmentation. This means that two words that should be separated by a whitespace instead appear together and vice versa.

Each token is then assigned a tag based on a pre-defined list. This tagging list contains variations of common identifiers such as block, town, roads etc. Any token not in this list is tagged as being the name of a place. The purpose of this step is to allow the algorithm to separate name tokens such as ‘Ahmed’ from identifier tokens such as ‘Block’ or ‘Road’. All numbers are tagged with the ‘Number’ tag. This step thus makes important abstractions. For example it will tag each number token (e.g. ‘55’) as a ‘Number’ and all name tokens as ‘Name’. It will also tag all variations of ‘Block’ as a ‘Block tag. Through this step the gains a partial semantic understanding of the address i.e. it now knows through the tag what each token might represent.

The tagged addresses are then run through a technique known as named entity recognition to extract chunks from each address. It uses a pre-defined grammar that specifies how different types of chunks may be arranged in the list of addreses. The pre-defined grammar would for example specify that a town chunk can be identified by one or more ‘Name’ tags followed by a ‘Block’ tag.

The extracted chunks are then fed into a data structure that keeps tracks of the number of times a chunk is encountered. It also builds a hierarchy of the chunks during the process. Addresses in general, are hierarchical in nature. This means that smaller units will be mentioned before larger units in the address, allowing us to extract a hierarchy of chunks. This hierarchy and count is the final output of this algorithm.

The algorithm assumes that key places within a feeder will be found repeatedly among the address list. This is necessary since not every address is parsed correctly by the algorithm due to variations in how addresses are specified. By assuming repeatability we are in effect assuming that for each address that is not correctly parsed there will be other addresses referring to the same place that will be parsed correctly. It was found that this assumption does not always hold. This is especially true in case where addresses refer to institutions such as airports, large hospitals or housing societies having a single or a few connections. These mostly have one or a small number of addresses thus appearing in the addresses with a low frequency. Since we did not manually go through the address list we discovered this through the reverse mapping process described later in the document.

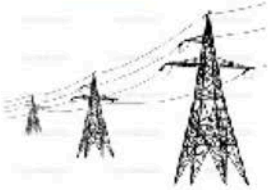
From Chunks to Mapping

Once chunks for each feeder have been identified the task of mapping these feeders remains. Chunks are name places that have appeared with a high frequency in the address data.

The primary idea is that each feeder’s spatial footprint would comprise the area of all the chunks associated with it. The process above was repeated for all 259 feeders of circle five.

Overall Results of the Geo-referencing Exercise

259



Feeders in circle V

500,000

20-A Abubakr Block
176 Shadman
Opp General Hospital
M Iqbal Ghurki pind
.....

Consumer addresses processed

2572

Abu Bakr Block

Garden Block

Shah Kamal

....

Address Chunks algorithimically extracted from 500,000 addresses

2015

Of 2572 address chunks successfully geo-referenced

Types of Chunks Identified and Geo-coding Strategy

Different types of chunks were identified by our algorithm and for each type of chunk a different strategy had to be used to map them.

Markets and Centers

Since each shop within a market or a commercial plaza is generally assigned its own electricity connection each market showed up in the addresses with a high frequency and identified as a separate chunk by our algorithm. We geo-coded markets and centers at the block level which means that assigned each market chunk to the block containing that market.

Villages

We mapped villages where we could find boundaries for them. Circle 5 contains many villages on the eastern outskirts of the city but boundaries do not exist for many of them. Village chunks were left un-mapped where we could not find their boundaries.

Colonies, Towns and Abadis

These chunks were straightforward to map.

Roads

The difficulties in geo-coding road chunks arise from the fact that although roads themselves are geo-coded in their entirety, address numbers along the road are not geo-coded. Therefore given an address on a road it is not possible to locate that address as a point.

Sometimes it will be known that a road is contained within a larger unit that we are able to geo-code. In these cases we have assigned the road chunk to this larger unit. For example *Zaildar* Road is entirely located within *Icchra* so we have assigned the chunk referring to *Zaildar* Road to *Icchra's* polygon. This approach introduces a level of coarseness in our output but is the best that can be done given the current state of geo-referenced databases for the city.

Sometimes road names accompanied by identifier a place identifier LIDHER B/ROAD LHR indicating the area along the road where these addresses are located. In this case we geo-coded this address in the polygon for LIDHER. In all other cases we have left road chunks unmapped.

Ambiguous Address Chunks

In some cases there is an inherent ambiguity in addresses. These addresses are defined in terms of landmarks and there is no indication of how far or how near the landmark they end or start. An example of this is addresses such as the one below identified by our algorithm:

OPP GENERAL HOSPITAL

Chunks such as the one above could not be mapped.

Other Geo-Coding Considerations

Reverse Geo-Coding

We found that there were important consumers such as hospitals that did not appear with a high frequency so therefore had not been identified as chunks by the algorithm. A reverse mapping exercise was carried out – by looking at a map of already mapped areas and identifying had not been associated with a feeder and then searching for these in the addresses to establish the feeder(s) serving them. This turned up a host of hospitals and institutions such as *Gulab Devi* hospital, Civil Services Academy and Walton Airport. Societies such as Model Town are mentioned only once in the address the reason being that LESCO supplies only one connection to the

Deciding on granularity

For some feeders geo-coordinates of a large number chunks identified by our algorithm were not available. In these cases we geo-coded the next larger unit, containing these chunks, whose coordinates were available.

For example chunks for Cantt consist largely of roads that are not possible to map. Therefore these road chunks were assigned as being within the polygon for Cantt. Another example of where we used this approach is while mapping *Icchra*. Our algorithm mined many different chunks within I$_{\text{cchra}}$, but these were all assigned to the larger polygon for *Icchra* due to the unavailability of the location of the exact chunks.

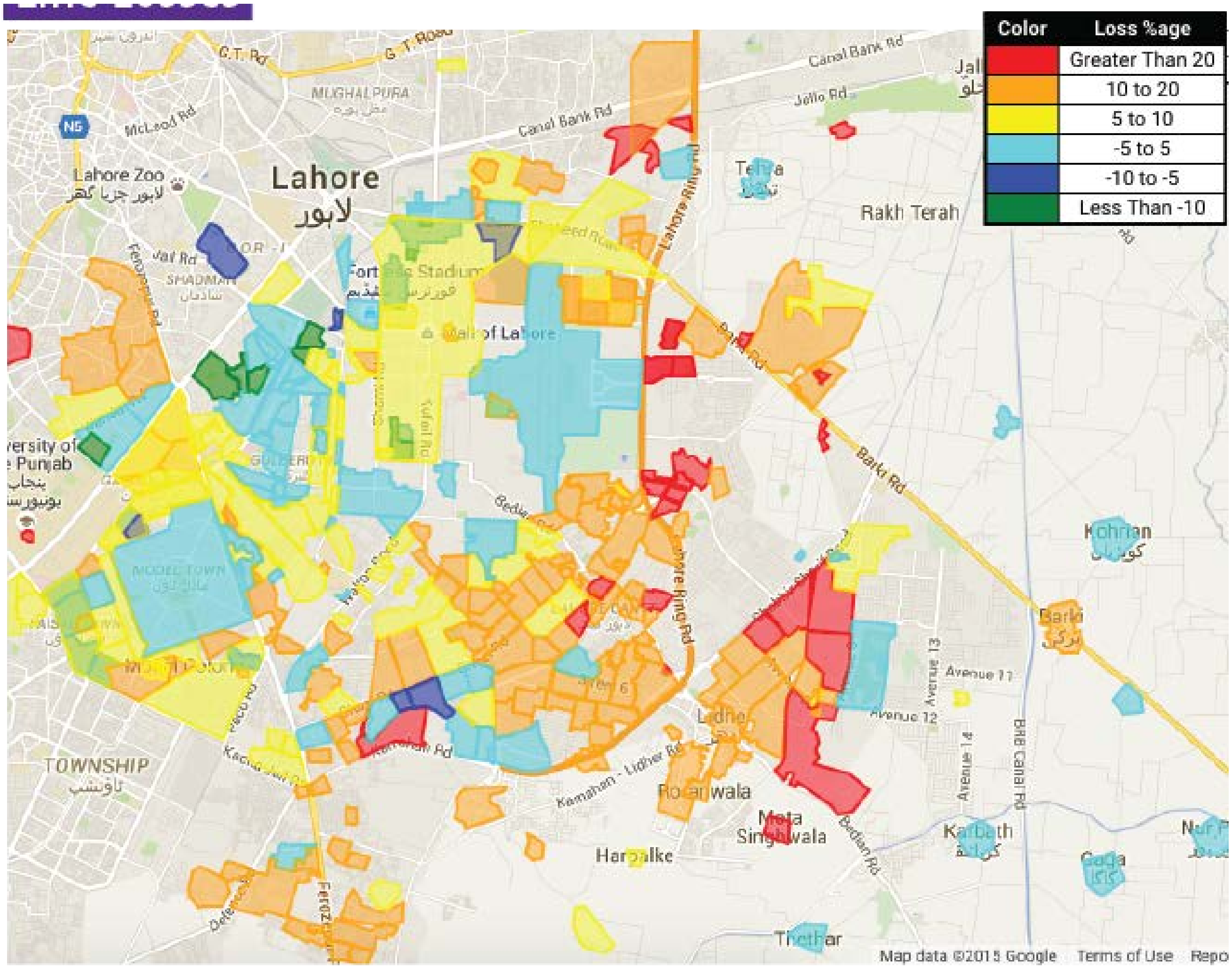
Applications

Spatial Representation of the Grid

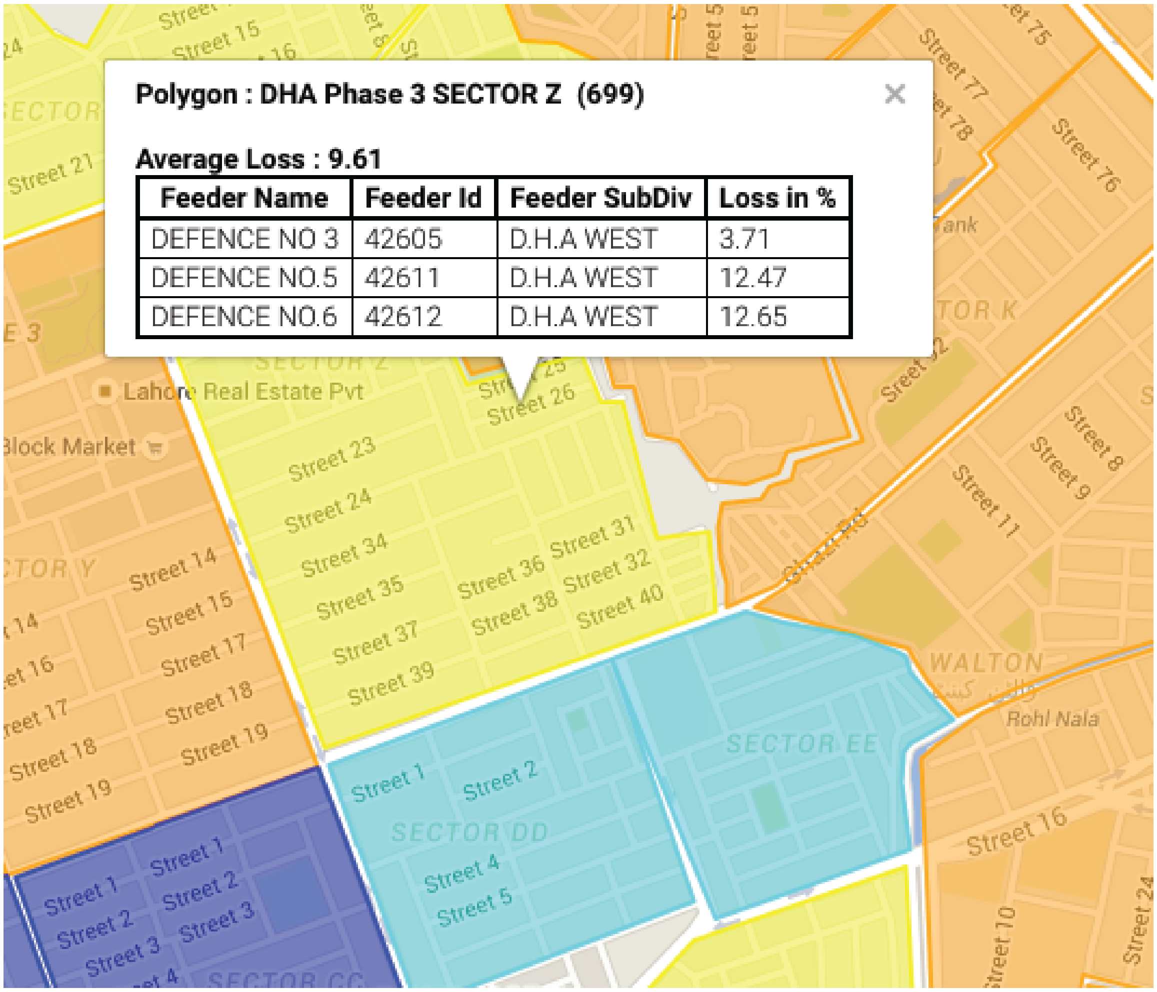
This is the first such representation that has been created for the grid in Pakistan. It can now be used to visualize variables such as consumption and line losses at the feeder level enabling a better understanding of the system.

One such pattern that immediately became apparent when we visualized feeder losses is that losses in the peripheral of the city are higher than those at the center. This is a fact that has been known to power officials and the fact that our visualization reflects this was of great satisfaction to us.

Line Losses



The visualization above shows losses for the various polygons mapped for circle V. Note that areas in the periphery generally have higher losses. In addition since each area can be served by more than one feeder the loss of percentage calculated for that area is the average of feeder losses serving that area.



Polygon : DHA Phase 3 SECTOR Z (699)

Average Loss : 9.61

Feeder Name	Feeder Id	Feeder SubDiv	Loss in %
DEFENCE NO 3	42605	D.H.A WEST	3.71
DEFENCE NO.5	42611	D.H.A WEST	12.47
DEFENCE NO.6	42612	D.H.A WEST	12.65

The above figure shows the losses for Sector Z in Defence phase 3. Three feeders, called Defence No.3, Defence No. 4 and Defence No. 5, serve the polygon. The loss for this polygon is calculated as the average of the losses of these three feeders.

Proxy for Socio-Economic Status

Socio economic datasets for Pakistan provide data at coarse levels of spatial granularity. Two datasets that are frequently used by economists and policy makers are the Multiple Indicator Cluster Survey and the Pakistan Standard Living Measurement (PSLM). MICS is representative at the tehsil level whereas the PSLM is representative at the district level. Researchers and policy makers frequently need socio economic indicators at a higher spatial granularity. For example a researcher may want to test the correlation between socio-economic status of neighborhoods and the level of crime in these neighborhoods.

It is well established that there is a strong correlation between power consumption and socio-economic status of a country. Extrapolating this relationship to smaller units such as blocks or towns within a city can enable this data to be used as a proxy for the socio-economic status within these areas. It can provide a powerful alternative to traditional sources of socio-economic data.

Addresses

The address processing algorithm is general in nature so that it can take as input any list of addresses and mine key areas for that list. It is likely that other organizations maintain such lists of addresses and the same process as that used in this project can be used to process and geo-locate their addresses.

Conclusion

Out of 500,000 addresses processed, 2572 chunks out of which 2015 were successfully geo-referenced. 102 Feeders were completely geo-coded whereas the remaining were geo-coded to varying degrees of completeness. The figure below shows the distribution of feeders by the percent of chunks successfully geo-coded.

Distribution of Feeders by percent of Chunks Successfully Geo-coded

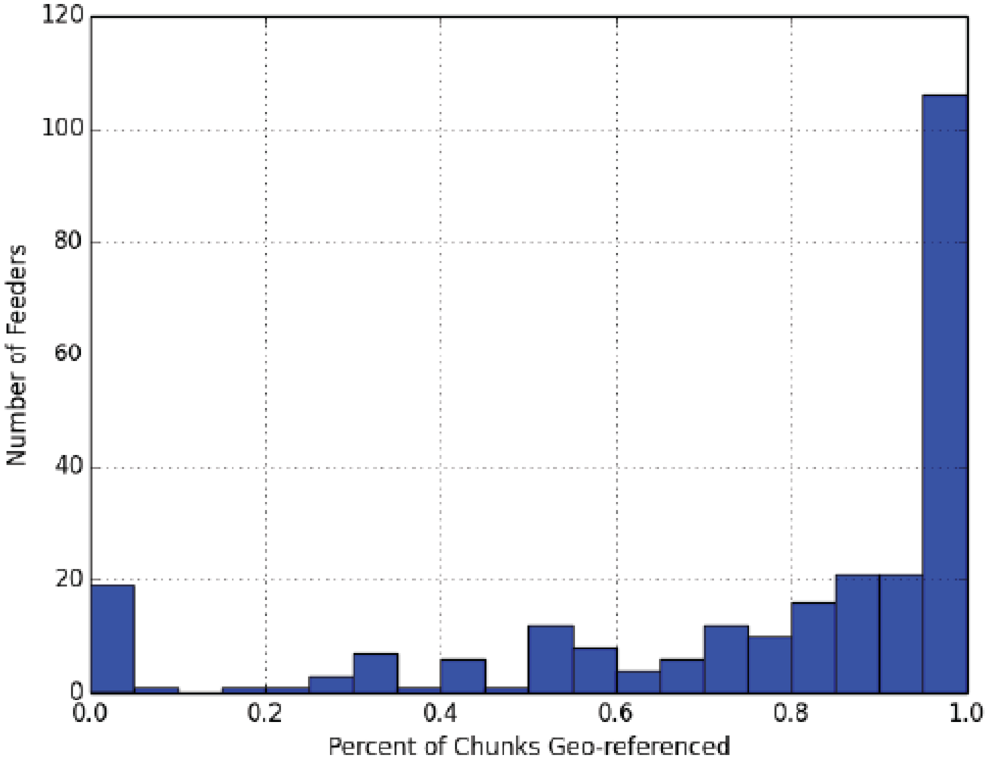


Fig 7 Distribution of feeders by the percent of chunks that were succeffully geo-`referenced.

Large feeders that could not be geo-referenced were mostly those that serve villages on the outskirts of the city. These are feeders such as 99103 (Hair), 98720 (Barki) and 98712 (Jahman). Since geo-coordinates of most of the areas in these feeders are not available a large percentage of these feeders remain unmapped. Therefore the unavailability of a complete geo-referenced database is a limitation of this approach.

Physical boundaries used were those obtained from Google or Wikimapia. It is conceivable that LESCO might have a different spatial definition of an area than these sources. We suspect this will be truer of older areas or abadis that evolve over time than newer ones such as DHA that are properly zoned.

Despite these limitations we believe that this approach provides a quick and low-cost solution to mapping the electricity system in urban areas of Pakistan. The effectiveness of this approach will increase with time as geo-referenced data for Pakistan becomes richer. Approaches such as crowd sourcing boundaries of chunks may be used in the meantime.

The International Growth Centre (IGC) aims to promote sustainable growth in developing countries by providing demand-led policy advice based on frontier research.

Find out more about
our work on our website
www.theigc.org

For media or communications
enquiries, please contact
mail@theigc.org

Subscribe to our newsletter
and topic updates
www.theigc.org/newsletter

Follow us on Twitter
@the_igc

Contact us
International Growth Centre,
London School of Economic
and Political Science,
Houghton Street,
London WC2A 2AE

IGC
International
Growth Centre

DIRECTED BY
LSE
UNIVERSITY OF OXFORD

FUNDED BY
UKaid
from the British people

Designed by soapbox.