TPI Technology for the People Initiative

Working Paper

Cell Phone Price as a Proxy for Socio-Economic Indicators

Syed Fahad Sultan (TPI)
Hamza Humayun (TPI)
Marium Afzal (The World Bank)
Umar Nadeem (The World Bank)
Sohaib Ahmad Khan, PhD (TPI and LUMS)

December 31, 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

Introduction

Socio economic data is not collected frequently in many developing countries. In Pakistan, the population census was last conducted more than fifteen years ago, despite the constitution requiring for it to be conducted every five years. This is primarily owing to the high cost and vested political factors entailing the existing practices of data collection on such a large scale [7]. On the other hand, the Information and Communication Technology (ICT) sector has grown enormously in Pakistan over the past couple of years. This growth has resulted in Pakistan having one of the highest tele-density levels in the region. As of 2012, there are around 118 million mobile subscribers in Pakistan [7].

High levels of tele-density yet lack of public socio economic datasets is a phenomenon not specific to Pakistan but is also prevalent in other developing countries [2]. However, therein also lies an opportunity. The low cost and real time data trail left behind by mobile users of a region can provide valuable insights into its socio economic state. Socio economic data is most useful when it is granular and up to date. Cellular usage data not only fits the bill on both these counts but since it is already being collected by telecom companies, there is very little cost involved in its availability.

Based on these intuitions, we set out to test our hypothesis and try to identify how mobile usage data of a region relates to its socio-economic indicators. For this study, we use eleven months of data of prepaid subscribers of district Jhelum in the Punjab province of Pakistan and then correlate it with variables in two publicly available socio-economic datasets: mauza census data and population census data. In addition to the mobile data variables obtained from the telecom company, we also analyze Phone Price, derived from handset names, a variable hitherto unexplored in the development literature.

This study is ongoing work which commenced two months ago after availability of cellular call records from out partner operator. In this report, we describe some initial results of our analysis of this dataset.

Related Work

Availability of cell phone usage datasets on a large scale is a recent phenomenon. A few interesting studies that try to understand how mobile data relates to socio-economic variables are discussed below. None of them however, study average phone price of a region as a variable.

One approach that has been taken by several researchers [2, 4, 3] is to collect socio-economic data by manually conducting interviews and relating the answers to the cellular usage behavior of the interviewees. Blumenstock et al. conducted such a study in Rwanda [4]. Their results revealed that significant differences exist between usage patterns of people from different socio-economic classes. Using a similar approach Kwon et al. [3] study acceptance of mobile phones and how it relates to demographics and socio-economic factors. A shortcoming of conducting interviews is that because of its high cost, the approach does not scale well. Our study addresses this problem by using preexisting datasets conducted by national census organizations. However, our study also does not study relationships at the level of individuals; rather we make inferences about the socio-economic indicators of a region covered by a cellular tower.

In terms of its overall theme, the work conducted by Soto and Frias-Martinez is close to our approach. In [1], they create a model to predict the socio-economic level of a region using 279 features from a subscriber's call detail records (non-aggregated metadata for individual calls). They report that their model can predict socio-economic level with up to 80% accuracy. However, employing such a large number of usage variables has its drawbacks. Not only may these variables not be available in most cases, but they also add complexity. Somewhat surprisingly, despite using so many variables, their work does not mention of price of the handset.

Data

In this section, a description of the datasets and the variables that were available to us for this study is given.

Cell Usage Data

For this study, we had access to eleven months of cell usage data of a major telecom provider in Pakistan, Mobilink, for the district of Jhelum. District Jhelum has an area of approximately 3,500 km^2, and a 2010 population estimate of approximately 1.2 million. About 72% of population is classified as rural, according to the 1998 population census. It was chosen because of geographic and socio-economic heterogeneity within the district and because it is one of the few districts for which for which hand-marked settlement boundaries were available. Mobilink has been the most popular cellular service provider in Pakistan[6].We did not have data for postpaid subscribers, but in Pakistan, prepaid subscribers make up for an overwhelmingly large percentage of total subscribers.

The data available for this study was not the actual Cellular Detail Records (CDRs). Instead, mobile usage variables were aggregated per day for a subscriber. So if a subscriber made use of any service on a day, a row is available in this dataset. All subsequent activity on that day for the same subscriber is aggregated in the same row. Separately, for each active day of each subscriber, we had a list of towers under which the subscriber performed some activity. This list of towers was not sorted in any order thus not allowing 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.

Deriving Phone Price from Call Records

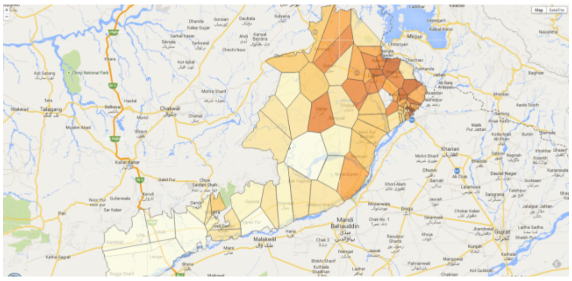


FIGURE 1. CHOROPLETH MAP OF AVERAGE PHONE PRICE, SHOWING SIGNIFICANT VARIATION WITHIN THE DISTRICT

In our dataset, one of the variables was the handset name of the subscriber against which activity for the day was recorded. For these handset names, we manually collected their existing market prices, by searching popular retail websites in Pakistan. At the end of the exercise, we managed to collect phone prices for around 85% of users. Handsets for which prices could not be found were almost all rare 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 Rs. 3000 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 like olx.com.pk and hafeezcenter.pk. For users against which activity was recorded from more than one phone over the eleven months, the average price of their phones was used.

Figure 1 shows a choropleth map of Average Phone Price over towers of Jhelum while Figures 2 and 3 shows distributions of the Phone Price variable, in Pakistani Rupee, over users and towers respectively.

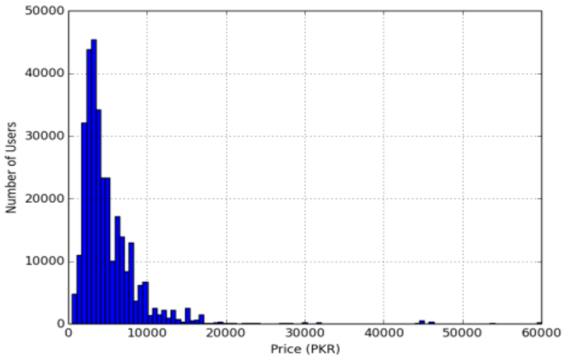


FIGURE 2. DISTRIBUTION OF PHONE PRICE OVER USERS

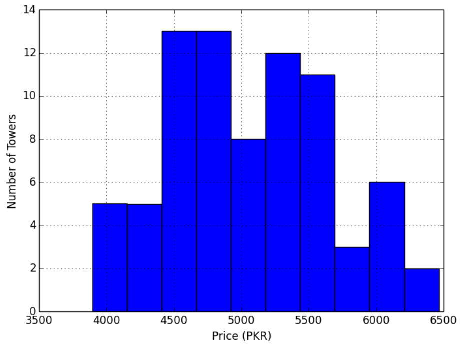


FIGURE 3. DISTRIBUTION OF AVERAGE PHONEPRICE OVER TOWERS

Census Data

We used two publicly available socio-economic census datasets in our study. The first is the Mauza Census of 2008, conducted by the Agricultural Census Organization (ACO), is an enumeration of key socio-economic indicators of each of 597 mauzas (revenue villages) in the district. Our second dataset is the

Population Census of 1998, conducted by the Population Census Organization (PCO). The 1998 census is the latest population census to be completed in Pakistan. From PCO and ACO datasets, we selected 16 variables as representative of the socio-economic state of a mauza. These along with their broad socio-economic theme and source have been listed in table 1.

TABLE 1 CENSUS VARIABLES SHORTLISTED FOR THIS STUDY

VARIABLE	SOURCE	THEME
Distance to College (Boys)	Mauza Census	Education
Distance to College (Girls)	Mauza Census	Education
Literacy Rate	Population Census	Education
Distance to Child/Mother Center	Mauza Census	Health
Distance to private MBBS doctor	Mauza Census	Health
Bricked streets*	Mauza Census	Infrastructure
Bricked Drains*	Mauza Census	Infrastructure
Construction Type of majority of houses*	Mauza Census	Infrastructure
Toilet facilities*	Mauza Census	Hygiene
Distance to police station/post	Mauza Census	Law & Order
Distance to private veterinary facility	Mauza Census	Livestock
Distance to govt. wheat/grain procurement center	Mauza Census	Agriculture
Distance to internet	Mauza Census	Communications
Availability of cable+	Mauza Census	Communications
Distance to commercial bank	Mauza Census	Financial
Distance to CNG or LPG	Mauza Census	Energy

*Smaller values indicate better availability and vice verse
+ Larger values indicate better availability and vice versa

There were a lot of common variables in ACO and PCO datasets. ACO variables were generally given preference owing to Mauza Census having been conducted much more recently. Also, most ACO variables were related to availability and access to different facilities. For each type of facility, there was a binary variable indicating presence/absence and a "distance from" variable which is zero in case the facility is present. So in short listing of the variables, we preferred distance variables on binary variables since they conveyed accessibility information in addition to availability. Also, the selection was performed making sure variables from each broad socio-economic theme was selected.

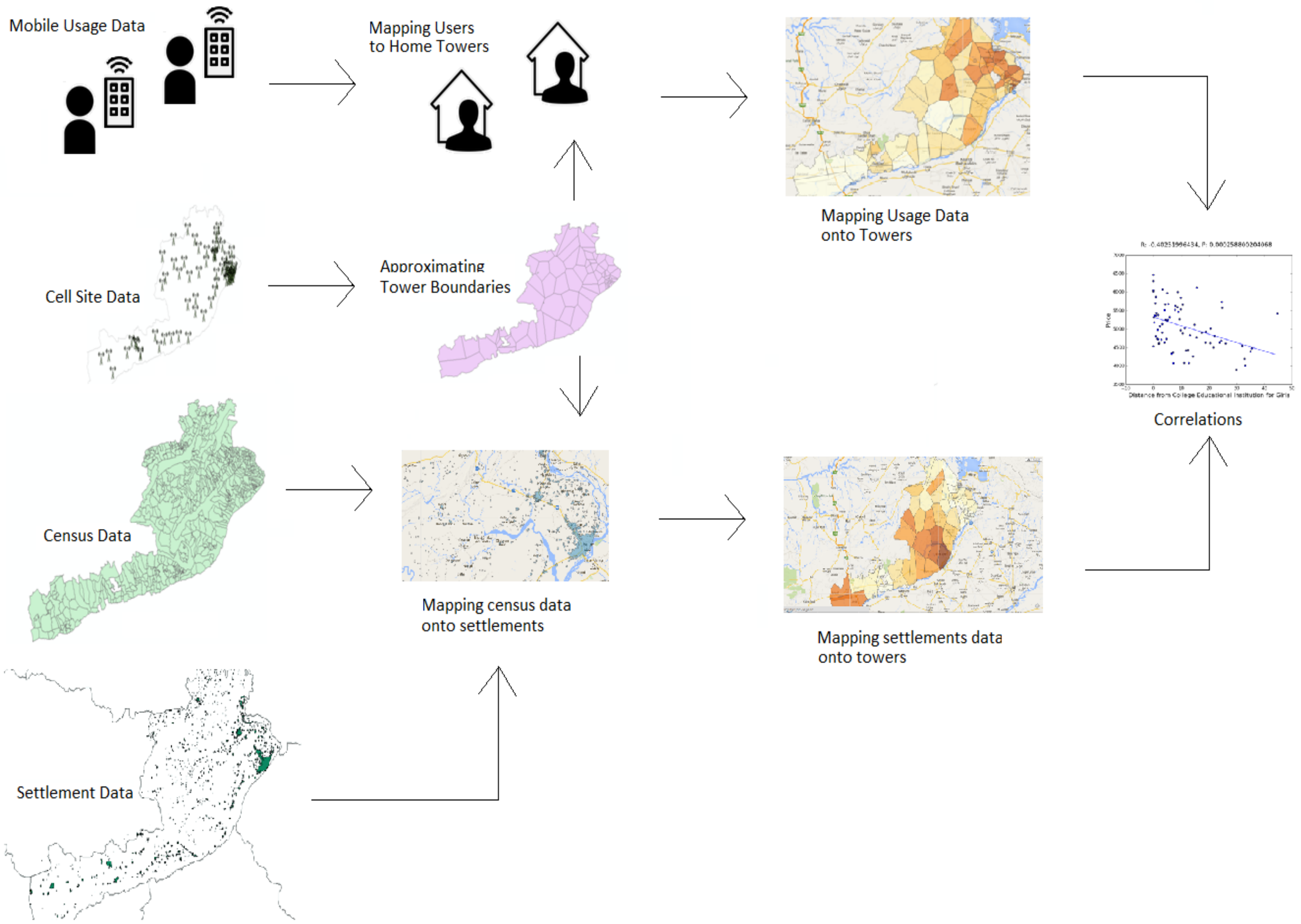

Mobile Usage Data
Mapping Users to Home Towers
Mapping Usage Data onto Towers
Cell Site Data
Approximating Tower Boundaries
R: -0.40251996434, P: 0.000258860264068
Price
Distance from College Educational Institution for Girls
Correlations
Census Data
Mapping census data onto settlements
Mapping settlements data onto towers
Settlement Data

FIGURE 4. FLOW DIAGRAM OF THE METHODOLOGY ADOPTED, SHOWING HOW THE DATASET IS PREPARED FOR COMPUTING CORRELATIONS

Furthermore, for each socio-economic theme, those variables were selected that exhibited a large spread or statistical dispersion, measured through inter-quartile range. A variable with a large spread indicates greater variation amongst mauzas of the district, which in turn points towards greater inequity with respect to that socio-economic variable.

Methodology

Mapping users to home towers

In the census data, socio-economic variables of a region reflect the condition of its residents. In order to correlate mobile usage dataset against it, we needed a method to approximate the place of residence of mobile users. In order to do so, we mapped users onto what we call their 'home towers', taken to be the tower under which their residence is located. Because of the unavailability of data within the day, the most frequent locations of subscribers during off hours could not be tracked. In light of this limitation, we instead estimated the home tower to be the tower under which the subscriber was spotted the most, overall in the eleven months.

Estimating cell site boundaries

In the mobile dataset, the towers were available to us as Latitude Longitude pairs. In order to get an approximation of their area of coverage, we used Voronoi algorithm [6]. The algorithm simply assigned region to each tower in such a manner that each point in the region is closer its assigned tower than any other tower.

Mapping census data to settlements

As the first step in mapping census data onto towers, we mapped census data onto settlements. We performed this mapping assuming flat distribution of census variable onto all of its settlements. Large settlements falling into multiple mauzas were split into smaller contiguous settlements.

Mapping data on settlements to towers

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

Correlations

Once census and mobile data variables 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 $|R| >0.3$
and $p < 0.01$.

Results

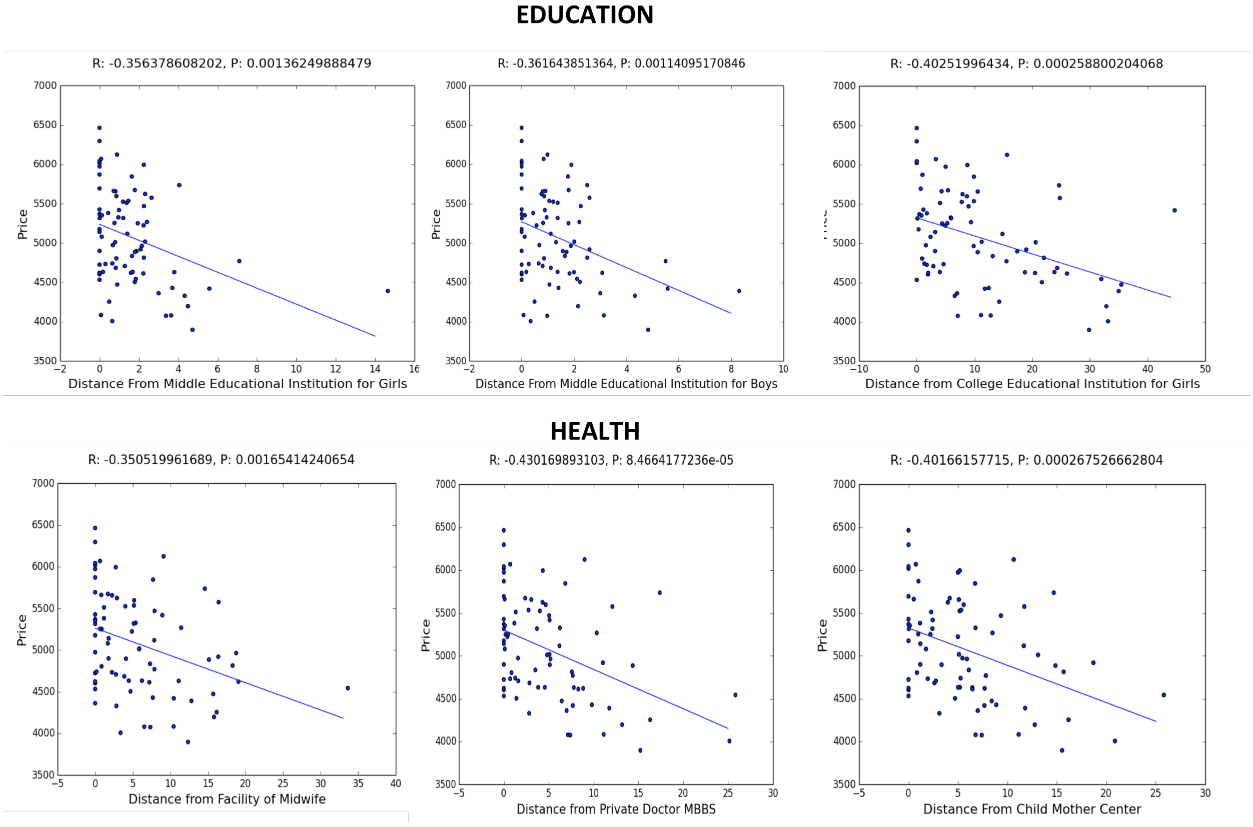
Average Phone price

Our results indicate that phone price correlates with a large number of socio-economic indicators. Though correlations with the individual variables is moderate, average phone price correlates broadly with a large number of socio-economic indicators and with indicators of varied socio-economic themes. Thus we believe that there is a clear relationship between the average phone price of a region and its socio-economic state.

It shows good correlation with accessibility parameters of several categories, like distance to college, police station, bank health care, veterinary facilities and grain procurement center. It also shows good correlation with infrastructure(bricked streets, bricked drains, construction type and toilet facilities) and communication (internet, cable) parameters. Interestingly, we did not find significant correlation between literacy parameters reported in PCO data and average phone price over a region. Table 2 lists socio-economic variables, their theme and the Pearson correlation coefficient R and p-value with phone price while Figure 5 shows scatter plots of the same correlations.

TABLE 2 CORRELATION COEFFICIENT (R) AND P-VALUE FOR CORRELATIONS 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	$8.7\times10^{-0.6}$
Distance to CNG or LPG	Energy	-0.43	$7.5e^{-0.5}$



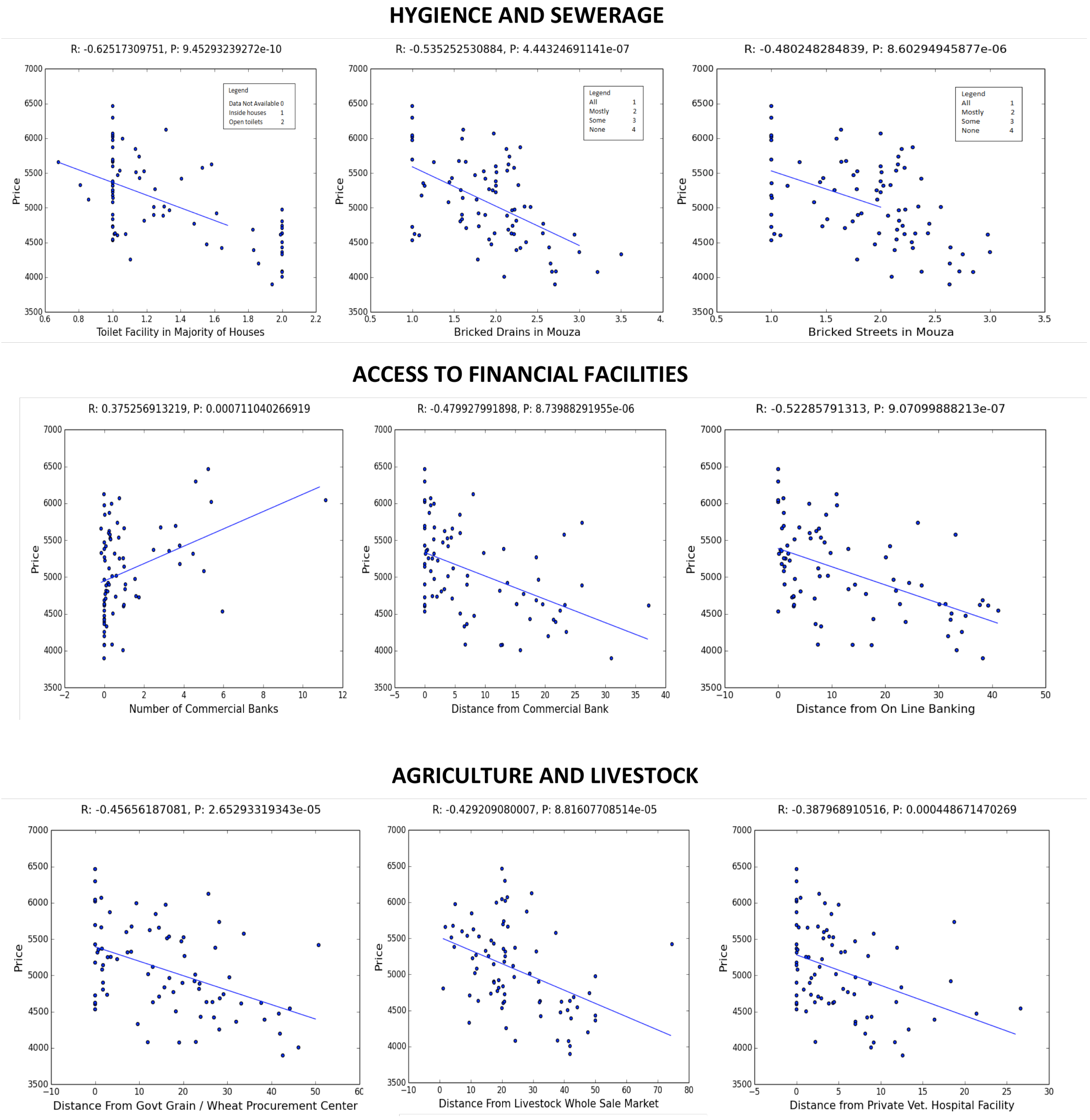


FIGURE 5 SCATTER PLOTS OF CORRELATIONS WITH AVERAGE PHONE PRICE

Mobile Usage Variables

Table 2 presents a correlation matrix with selected census variables as rows and mobile data variables as columns. The mobile data variables have been ordered by number of correlations with census variables.

TABLE 3 CORRELATIONS OF MOBILE USAGE VARIABLES WITH CENSUS VARIABLES

	Price	SMS Revenue	SMS Outgoing	Closing Balance	Calls Revenue	SMS Incoming	Total Revenue	Recharge Amount	Calls Minutes.	Number of outgoing calls	Number of Recharges	GPRS Revenue	Minutes of incoming calls	Number of incoming calls
*Boys College	✓	✓	✓	✓	✓	✓	✓	✓	✓					
*Girls College	✓	✓	✓	✓										
Literacy Rate														
*Child Mother Center	✓	✓												
*Private Dr. MBBS	✓	✓												
Bricked Streets	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Bricked Drains	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓		
Construction Type	✓	✓	✓	✓	✓	✓	✓	✓		✓				
Toilet Facilities	✓	✓												
*Police Station/ Post Office	✓	✓	✓											
*Private Veterinary Hospital	✓													
*Govt. Grain/ Wheat Procurement Center	✓	✓												
Cable	✓	✓	✓	✓	✓	✓	✓	✓						
* Internet	✓													
*Commercial Bank	✓	✓		✓	✓									
*CNG or LPG	✓	✓												

* Distance from facility

Red tick marks indicate negative correlations while green ones indicate positive correlations. As mentioned earlier, two variables are said to be correlated if the absolute value of the Pearson correlation coefficient $|R|$ is greater than 0.3 and if the p-value is less than 0.01.

An interesting result is that literacy is the only variable that does not correlate well with any mobile data variable. Our experiments indicate that mobile data variables do not correlate well with other Population Census variables, not listed here. This might be owing to the Population Census data having been collected over 15 years ago.

Also note how variables related to outgoing activity correlate better with socio-economic variables as compared to incoming activity. This is understandable and was to be predicted since outgoing activity results in expenditure of capital. Furthermore, interestingly, expenditure resulting from SMS activity is much more reflective of the socio-economic state of a region as compared to expenditure from calls activity.

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 several of its socio-economic indicators. We also show that variables related to expenditure 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 is ongoing work and this paper presents some initial results. 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.

Acknowledgment

We are grateful to Mobilink for providing this dataset as part of their corporate social responsibility.

References

[1] V. Frias-Martinez and J.Virseda. On the relationship between socio-economic factors and cell phone usage. *In Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*. ICTD '12, pages 76-85, New York, NY, USA, 2012. ACM

[2] Smith C., A. Mashhadi and L. Capra (2013) *Ubiquitous Sensing for Mapping Poverty in Developing Countries*. Presented at NetMob 2013 for the D4D Challenge.

[3] H. Kwon and L. Chidambaram. A test of the technology acceptance model: The case of cellular telephone adoption. *In Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000.

[4] J. Blumenstock and N. Eagle. Mobile divides: Gender, socioeconomic status, and mobile phone use in rwanda. *In Proceedings of the 4th International Conference on Information and Communication Technologies and Development*, 2010.

[5] G. Voronoi. Nouvelles applications des parametrescontinus a la theorie des formesquadratiques. *Journal fur die Reine und AngewiandteMathematik*, 133:97-178, 1907.

[6] Telecom Indicators. Retrieved December 20, 2014 from http://www.pta.gov.pk/index.php?Itemid=599

[7] Abdul Manan, 2014. Long Delayed Count. Retrieved December 20, 2014 from http://tribune.com.pk/story/671584/long-delayed-count-govt-drops-census-plan-for-fear-of-backlash/