The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-4/W5-2022
7th International Conference on Smart Data and Smart Cities (SDSC), 19–21 October 2022, Sydney, Australia

LAST MILE LOGISTICS: IMPACT OF UNSTRUCTURED ADDRESSES ON DELIVERY TIMES

M. Abdul Rahman*, M. Aamir Basheer, Z . Khalid, M. Tahir , M. Uppal

Department of Electrical Engineering, Lahore University of Management Sciences, Lahore 54792, Pakistan
(muhammad.rahman, muhammad.basheer, zubair.khalid, tahir, momin.uppal)@lums.edu.pk

Commission IV, WG IV/9

KEY WORDS: Urban logistics, Optimization, Last mile delivery, Geo-coordinates, Levenshtein distance, Address standardization

ABSTRACT:

The e-commerce industry has seen significant growth over the past few years. One significant issue that has sprung up as a result of this growth is unstructured addresses during last mile delivery. These ambiguous addresses are an established issue, particularly in developing countries like Pakistan. They are difficult to read and locate by last mile delivery riders thereby increasing delivery times and cost, negatively impacting the business of the company. Increased delivery times are also detrimental to the environment. In this paper, we aim to quantify the effects of unstructured addresses on last mile logistics. Many attempts have been made to standardise addresses to tackle this problem. Deep learning based approaches using recurrent neural networks (RNN) as well as probabilistic approaches using hidden Markov models (HMM) have been used. However, the main downside to these approaches are the underlying variation in address schemes in housing societies. We present an end to end rule based pipeline using Levenshtein distance (LD) and regular expressions (RegEx) rules which takes those unstructured addresses and outputs their structured forms along with their Geo-coordinates. The pipeline also returns the optimized route to minimize the last mile distance traveled.

1. INTRODUCTION

In recent years, the e-commerce industry has experienced significant growth worldwide. The COVID-19 pandemic has further accelerated these trends by altering the traditional shopping patterns of individuals which is seen in a better financial performance of logistics firms in the year 2020 (Atayah, 2021). This has enabled many firms, particularly micro, small and medium-sized enterprises, to stay afloat. For example, the e-commerce industry accounted for $3.351 trillion worldwide in sales for the year 2019 and is expected to reach 6.169 trillion (22.3% share of total retail sales) (Leblow, 2021). In Pakistan, we have witnessed a similar growth, that is, the e-commerce industry is worth $548.89 million in 2021 with a growth of 35% in value in the first quarter of 2021 (Shezad, 2021). This growth has also led to large losses due to poorly structured addresses. It is estimated that the impact of ambiguous addresses cost the logistics industry $6 - $8 billion in India alone (Bhattacharya, 2018). This cost is mainly due to inefficient route and delivery planning, last mile delays and lost productivity because of such addresses. The cost savings from standardized addresses for only the e-commerce sector is estimated to be about a $100 million (Bhattacharya, 2018). In this work, we aim to quantify the impact of unstructured or ambiguous addresses on last-mile logistics in urban areas. In our analysis, we use the data provided by the logistics company with operations across the country. In Pakistan, house-level Geo-coordinates are not available even for legible addresses on well-known routing applications which makes the address even more difficult to locate. Due to the ambiguity in locating an address, last-mile delivery riders encounter an increase in the distance traveled and time required for the delivery. In this paper, we propose a data-driven modeling framework to quantify the impact of ambiguous or unstructured addresses on last-mile delivery times. A summary of the framework is shown in Fig 1.

To address this problem, deep learning based approaches have been used extensively in the literature. Lu et al used a seq2seq recuurent neural network (RNN) based encoder-decoder architecture with bahdanau attention. They used the gated recurrent unit (GRU) in the RNN block (Lu, 2019). Probabilistic approaches have been used as well. Christen et al trained an hidden Markov model (HMM) to standardise addresses for a Geo-Coding system using the geocoded national address file (GNAF) data (Christen, 2004). Kaleem et al also train a HMM based model to standardize addresses (Kaleem, 2011). The main downside to these approaches is the underlying variation in address schemes in housing societies. Thus, training any model becomes tedious since effectively, a different model would have to be trained for each society. Therefore, in this paper, we develop a rule-based algorithm that takes ambiguous and unstructured addresses as input and uses the digital maps along with the data of the past deliveries to determine the structured address and corresponding Geo-coordinates.

2. DATA SOURCES

The data-set used in this paper was provided by Muller and Phips (M&P) courier company that consists of 5816 deliveries made during the second week of September 2021. For each delivery, we use the following information: timestamp, latitude, longitude and address. We also use the road network and the Geo-tagged digital maps of different suburbs of Lahore. The most number of 'unique Geo-Coordinate' deliveries was 134 being performed on 13-09-2021. By 'unique Geo-Coordinate' deliveries, we mean those deliveries that had distinct pin locations. For example, a rider may have delivered 10 parcels on any given pin location which appear as 10 separate entries in the data-set. However, since the location would be the same for all these deliveries, they will be counted as a single 'unique Geo-Coordinate'.

For our evaluation, we have taken 13-09-2021 as a case study since it had the largest number of 'unique Geo-Coordinate' deliveries. Fig. 2 shows the spread of deliveries according to rider

* Corresponding author

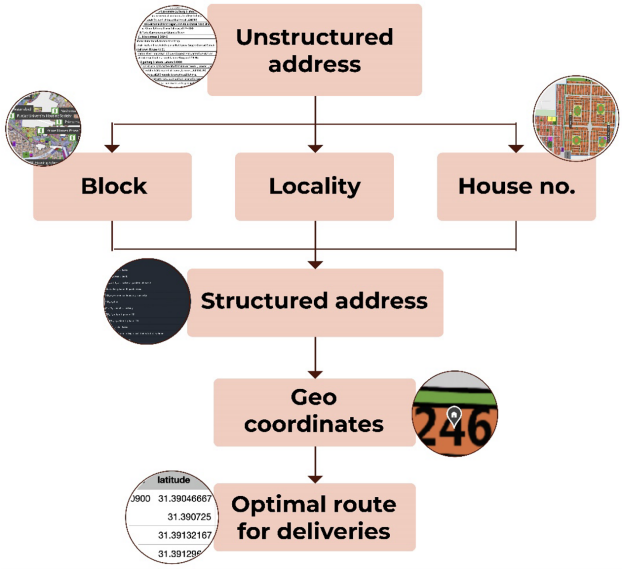

Figure 1. Representation of end to end pipeline for optimized route planning

number. The clusters in Fig. 2 clearly show that riders have particular zones in which they operate. They travel and perform deliveries in those areas only. As a result, the distance covered between two particular deliveries is relatively short. Evidence for short distances traveled by the delivery riders can also be seen in Fig. 3. The distance between deliveries for each rider does not have a large spread around their relatively low means. The maximum average distance covered by a rider is 1425 meters while the minimum is 490 meters for the deliveries shown in Fig. 4.

3. METHODOLOGY

3.1 Impact of Unstructured Addresses

For each origin and destination (OD) pair in the data, we use the QNEAT3 plugin in QGIS software and road network to generate the origin-destination (OD) matrix containing the shortest path distance for each origin and destination in the delivery data. The plugin uses the dijkstra's algorithm to determine the shortest distance between an OD pair using the road network. Table 1 shows the partial OD matrix. The road network shown in Fig. 5 has been acquired by the national transport research centre (NTRC). It contains the following classes of roads: primary, secondary, local, motorway, metro and highway. The network was edited to keep only primary, secondary, local and highways since it was assumed that a delivery rider is on a motorbike. Given the road structure in Lahore, motorbikes can only travel on these edited roads.

The network also contains a directions field with North Bound (NB), South Bound (SB), East Bound (EB), West Bound (WB) and None as one of the values for each respective road. NB and EB were assigned the forward direction of 1, SB and WB were assigned the backward direction of 2 whereas 0 was assigned for both directions for all the remaining roads with None value. These values are determined after analysing the ground truth road directions. The values are fed into the QNEAT3 plugin for direction based distances. Using the distance d (in meters) and the path information, we determine the travel time, denoted by t_1, for each origin and destination using the following equation

Origin ID	Destination ID	Distance (m)
0	1	1933.86
0	2	2453.12
0	3	2991.10
1	0	1933.86

Table 1. Distance in OD matrix

where f_s is average free flow speed (km/h) of the delivery rider, as follows

$$t_1 = \frac{3.6d}{f_s}, \tag{1}$$

We compute t_1 for the following values of f_s: 10, 15, 20, 25, 30, 35, 45 and 50 km/h. Using this information, we model the time (stochastic) taken by the delivery rider in finding the address as follows

$$z = t_2 - t_1 - y, \tag{2}$$

where t_2 is the actual travel time of the rider and y is a random variable which takes into account traffic congestion on a route, the waiting time at the delivery location, weather on the day and the variation in the route taken by the rider. We model the contributions of traffic congestion on a route towards y as

$$y_1 = 3.6 \times (\frac{1}{c_s} - \frac{1}{f_s}), \tag{3}$$

where the congestion speed c_s is determined as follows

$$c_s = \frac{t_{5\text{am}}\, f_s}{t_{\text{real}}}. \tag{4}$$

We determine $t_{5\text{am}}$ and t_{real} using the Google routes API as travel time at 5 am (reference time, assuming no congestion) and at the actual timestamp of the delivery respectively.

We model the contributions of waiting time i.e., time taken by a rider to mark the delivery in the system after having reached the door as follows

$$y_2 = \frac{\sum_{n=1}^{N} (t_2 - t_1)}{N_R}, \tag{5}$$

where the distance between t_2 delivery and t_1 delivery is less than 50 meters and N_R is the number of deliveries performed by the rider R. Weather effects are ignored owing to the prevalent conditions on that day. Similarly, uncertainty in the route is negligible since the riders generally follow the shortest route because of their knowledge of the road network in an area.

3.2 Address to Geo-Coordinates Converter

3.2.1 Levenshtein Distance The LD algorithm which was developed by Levenshetein (Levenshtein, 1966) takes two strings, namely the input string and the target string. The algorithm changes the input to the target with the minimum number of edits or operations on characters. These operations are insertion, deletion or substitution of characters, each operation having its own score. The path or the solution requiring the minimum number of operations is the LD for two particular strings (Bard,

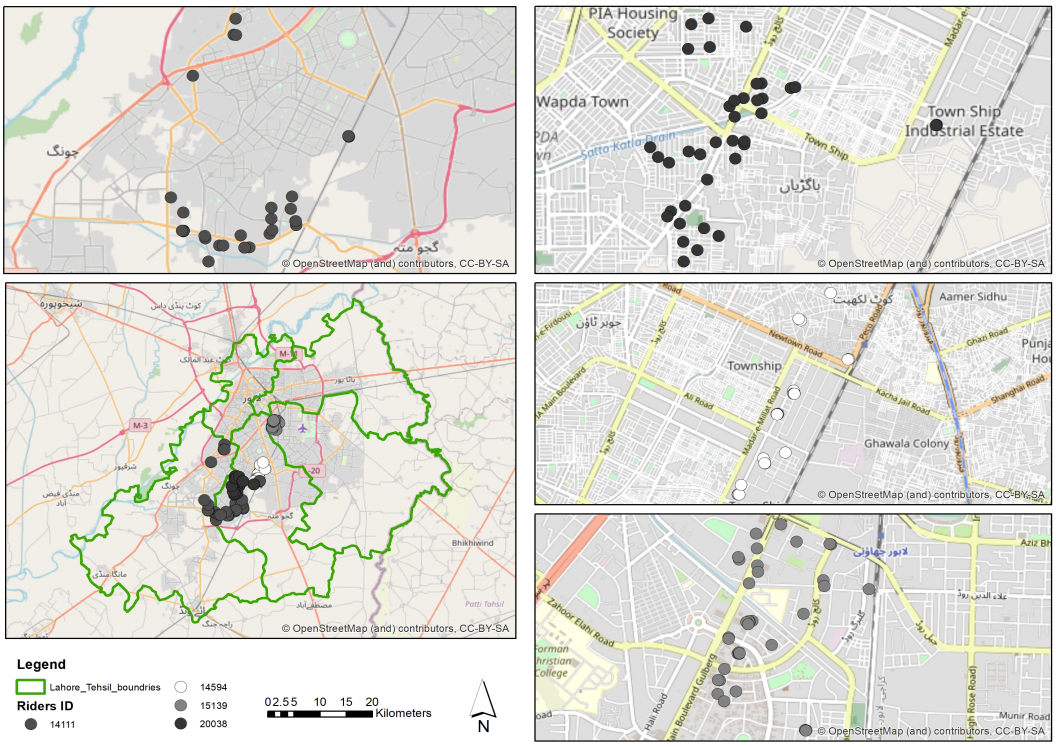

Figure 2. Spread of deliveries on 2021-09-13

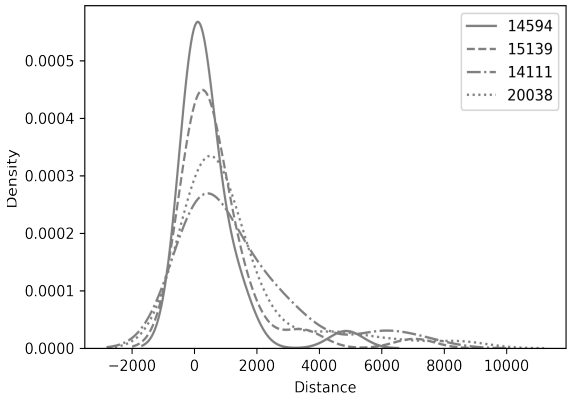

Figure 3. Kernel density estimate of deliveries (m)

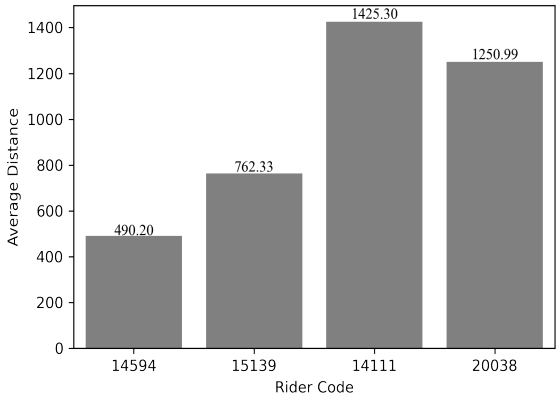

Figure 4. Average distance traveled by rider (m)

2006). Equation 6 represents the LD algorithm where $I(c)$ represents the insertion of a character, $D(c)$ is deletion and $S(c)$ is substitution.

$$\text{LD}(\text{Input}, \text{Target}) = \text{minimize}[I(c) + D(c) + S(c)]. \quad (6)$$

3.2.2 Converter Because of the impact of unstructured addresses on delivery times, we have developed a rule based system that converts an unstructured, messy address to its corresponding structured form. Additionally, Geo-Coordinates of the address is an output of the system. The system relies mainly on the LD and regular expressions (RegEx) rules to determine which society the address belongs to, followed by searching the block in the address using a predetermined list. The house number is the first element to be searched. The tool developed here can help courier services working in Lahore, Pakistan to convert unstructured addresses into structured ones. This will not only be beneficial for the courier services but will also help to mitigate the external factors associated with the last-mile logistics.

Initially, a deep learning based RNN model was used to tackle the problem. However, due to the amount of human resource required to label the training data and the variation in addresses writing schemes, a rule based approach has been adopted. In the rule based system, the address is first converted to lower case. Then, all the punctuation marks are removed and roman numerals are converted to integers. Thereafter, house number is determined for which the RegEx library of python is used. The maximum allowed house number is 10000, meaning if an integer is determined in an address that is greater than 10000,

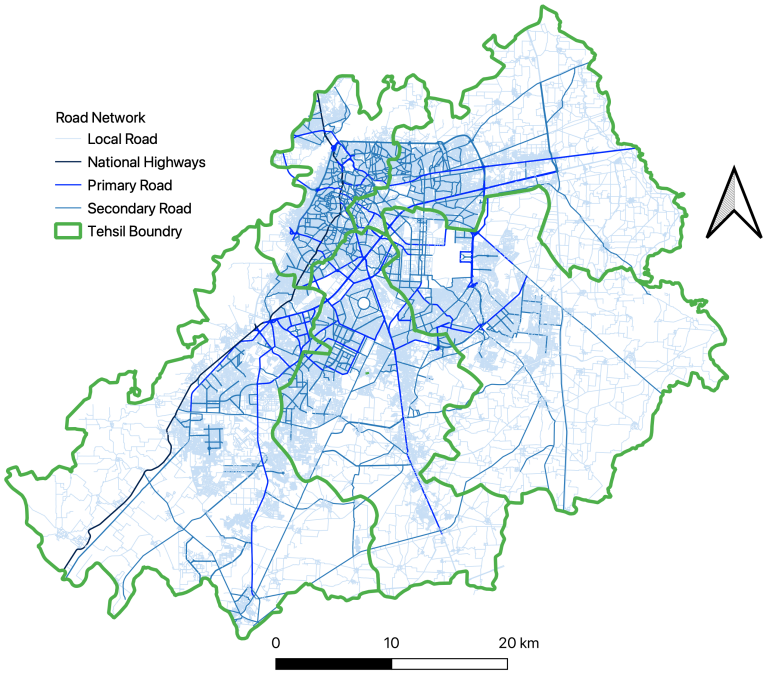

Figure 5. NTRC road network of Lahore

'-1' is returned as a flag. This threshold is used to avoid returning any spurious numbers usually found in addresses as house numbers. Secondly, the society of the address is determined for which LD is used. A society may have up to three parts which are Residence, Locality and Additional Identifiers (AID). The address is checked against a predetermined list of societies. All the elements in the list are checked in the address and their LD was determined. The society with the lowest possible LD is returned. Take for example DHA in Lahore. DHA is considered to be a residence with many localities, for example, phase 1, phase 8 etc. Phase 8 in turn has two AIDs namely, air avenue and park view. So an address may have "DHA Phase 8 Air Avenue" as a society. In case AIDs are not found, then "DHA Phase 8" may be returned. Finally, only "DHA" may be returned in some cases if only the residence is found.

Thirdly, the block of the society is derived using a predetermined list of blocks. LD is used here as well. Blocks in some cases may be simple, such as 'a' or 'b' and in some cases more complex such as 'sector b1 block 4' etc. Here as well, some addresses may only contain 'sector b1' in which case only 'sector b1' will be returned. If 'block 4' is only present, 'block 4' will only be returned, even though the predetermined list had 'sector b1 block 4' as a block. In all cases, if an element is not found, '-1' is returned as a placeholder. In addition to '-1', the block element has one more placeholder which is '-2'. This is for all those societies that do not have blocks in them. The Geo-Coordinates of the address are determined using an existing database. The predicted address has an associated Geo-Coordinate flag from the following list; 'accurate up-to house', 'accurate up-to block', 'accurate up-to locality', 'accurate up-to house, block compromised', 'accurate up-to block, block compromised', 'could not find the Geo-Coordinates, society not mentioned', 'could not find the Geo Coordinates, block was not unique' and finally 'could not find the Geo-Coordinates'. The first three flags are self explanatory. The fourth flag, 'accurate up-to house, block compromised' is for all those societies which have blocks with multiple elements but some element was missing in the address. Take for example township society. Township has sectors which in themselves have multiple blocks, one such example is "sector b1 block 4". So if an address is "house 29 sector b1 township", that is the second part of the block is not mentioned , the output will be "29, b1, township" and the corresponding flag will be the fourth one. This means that the system did find a house 29 in sector b1 of township but since the second part is not known, the overall block is compromised since the house could be found in either "sector b1 block 4" or "sector b1 block 1" etc.

The fifth flag is similar to the fourth one, the only difference being the house number mentioned in the predicted address could not be found in the database. The sixth one is for the explicit case in which the society is '-1'. If an address has the flag 'could not find the Geo Coordinates, block was not unique', it means that the predicted society had multiple blocks with the same name, but other parts of the society were not mentioned in the address. Take for example the address "house 23 block n dha". The predicted address would be '29, n, dha' but since dha has multiple phases with the same block that is, phase 1 also has a block n as well phase 8 and many other phases as well, the block was not unique of the society. If all else fails, the flag 'could not find the Geo-Coordinates' is returned. The LD in cases of both the blocks and societies varies depending on the length of the input string i.e the string in the predetermined list. If the string is less than or equal to three characters, the minimum LD is zero else it is two. If the input string matches with some part of the address but the LD is greater than the minimum standard LD previously determined, '-1' is returned as a placeholder. New societies are added to the list using the following method. The word or gram in the input address used to match an address to a particular society is returned, as well as their corresponding LDs. The gram is manually checked for any

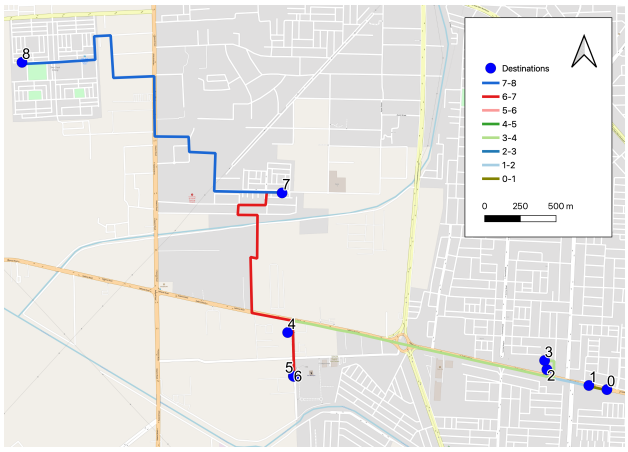

Figure 6. Snapshot of suggested route.

Distance (m)	Latitude	Longitude
8078.29969	31.39047	74.26316
	31.39073	74.26183
	31.39173	74.25874
	31.39230	74.25857
	31.39406	74.23964
	31.39132	74.24006
	31.39130	74.24009
	31.40284	74.23923
	31.41105	74.22007

Table 2. Suggested route

variations with the word in the predetermined list. If the word is a variation of an existing society name, the additional name is added to the list of societies along with the original one, otherwise, a new society is incorporated into the list. Along with the Geo-Coordinates and structured forms of addresses, our system also outputs a pin location of each address for which latitude and longitude pairs are found. This will help with visualization and route planning of the deliveries to be made.

4. APPLICATIONS AND ANALYSIS

Using the same data, we also undertook the task of developing a system that suggests an optimized route plan for a rider using Geo-Coordinates of the deliveries to be performed. QGIS aided by a python script was used for this task. The OD matrix again is determined using the QNEAT3 plugin in QGIS with the NTRC road network. For this purpose, we use the well established traveling salesman problem from operations research. The first Geo-Coordinate provided is assumed to be the starting point, from which the next shortest point is suggested, followed by the next and so on. Using our system, we also suggest the shortest route as shown in Table 2. We also provide a snapshot on a map of the suggested route that should be used while traversing shown in Fig. 6. The expected value of the time to find an address for each delivery due to ambiguous unstructured addresses varies from 117.42 to 219.32 seconds, depending on the free flow speed as illustrated in Fig. 7. This additional time leads to economic losses by virtue of the extra fuel consumed and distance traveled. It has a pernicious effect on the already degrading environment, especially with regard to air pollution.

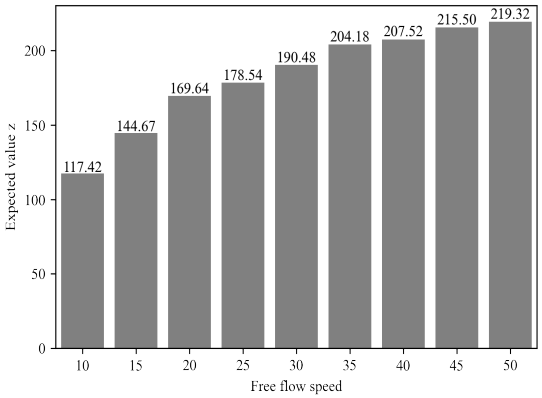

Figure 7. Expected value of the time to find an address for each delivery due to the ambiguous unstructured address against the free flow speed.

5. CONCLUSION

In our work, we have quantified the negative effects of unstructured addresses on last mile delivery times. The extra time spent for a delivery rider to find an address leads to additional costs for the logistics company. It also has a detrimental effect on the environment owing to the extra fumes of greenhouse gasses produced by vehicles of the riders. Having used these effects as a motivator, we have described our rule based system which takes unstructured addresses and outputs their structured forms including their respective Geo-Coordinates. A rule based system was preferred to a learning based or a probabilistic one because of the variation in address schemes from one housing society to another. Training based approaches would have required a significant amount of data plus the huge amount of manual labour to annotate that data. Using the Geo-Coordinates, we also suggest an optimized route for last mile deliveries. We hope this work would inspire industry-standard products which would help alleviate operational concerns of the logistics industry in Pakistan and provide a breathing space for the already degrading environment.

6. FUTURE WORK

We also plan to work on a number of threads in the future, one of which is Hub Optimization: determining the optimal place for hubs in the city which would lower the overall distance traveled by a rider. As of today, M&P has three hubs in Lahore to which each rider is allocated. Each rider collects their parcels from one of these and performs deliveries in their respective zones, these hubs however are not optimally placed. Another possible thread is Rider Allocation: determining the optimal number of riders to be allocated to a region based on the number of deliveries in that region to reduce operations and management costs.

ACKNOWLEDGEMENT

We acknowledge the support provided by M&P logistics and National Transport Research Centre (NTRC) for their active collaboration and for providing us with the necessary data sets used in this research work.

REFERENCES

Atayah, O., 2021. Impact of COVID-19 on financial performance of logistics firms: evidence from G-20 countries. *Journal of Global Operations and Strategic Sourcing*.

Bard, G., 2006. Spelling-error tolerant, order-independent passphrases via the Damerau-Levenshtein string-edit distance metric. *Cryptology ePrint Archive*.

Bhattacharya, S., 2018. Economic Impact of Discoverability of Localities and Addresses in India. *arXiv preprint arXiv:1802.04625*.

Christen, P., 2004. A Probabilistic Geocoding System based on a National Address File. *Proceedings of the 3rd Australasian Data Mining Conference, Cairns*.

Kaleem, A., 2011. Address Standardization using Supervised Machine Learning. *International Conference on Computer Communication and Management*.

Leblow, S., 2021. Worldwide ecommerce continues double-digit growth following pandemic push to online. https://www.emarketer.com/content/worldwide-ecommerce-continues-double-digit-growth-following-pandemic-push-online.

Levenshtein, V., 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady. Vol. 10. No. 8.*

Lu, Y., 2019. Chinese Address Standardization Based on seq2seq Model. *Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems*.

Shezad, A., 2021. Pakistan e-commerce platform Daraz aims to beef up as Amazon eyes market. https://www.reuters.com/business/retail-consumer/pakistan-e-commerce-platform-daraz-aims-beef-up-amazon-eyes-market-2021-11-25/.