Applying Operations Research Principles to Improve Geo-database Quality

Lean Principles and Value-Stream Mapping as Tools for Improving Geocoding Success

2014 ESRI International User's Conference

Data Compilation Case Studies: QA/QC

July 17, 2014

San Diego, CA

Anirudh Kannan Vinkayaram[1], Jack Baker[1], Ron Lumia[2], Ganesh Balaksharan[3,4], Nathan Crouse[1]

[1] Geospatial and Population Studies, University of New Mexico

[2] Mechanical Engineering, University of New Mexico

[3] Electrical and Computer Engineering, University of New Mexico

[4] Center for High Tech Materials, University of New Mexico

INTRODUCTION

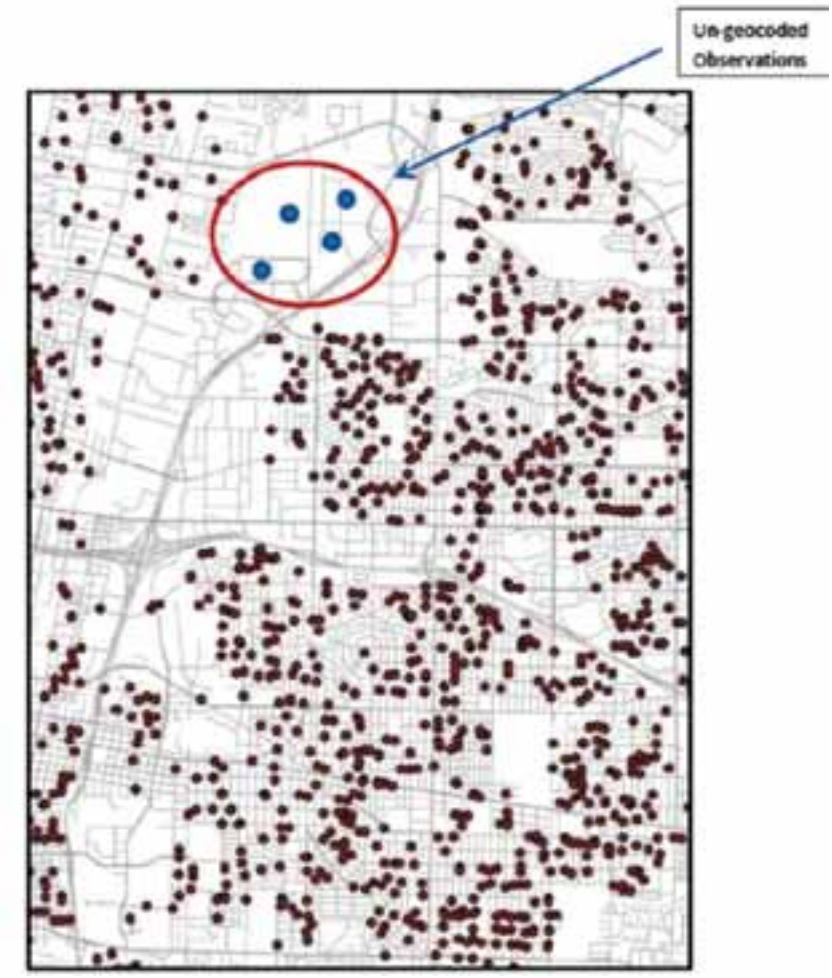

Figure 1. Incomplete Geocoding

Incomplete geocoding comprises an ever-present challenge to social science research that requires the placement of event data on maps at fine geographic scales as in Figure 1. In the case of small-area demographic estimates, the issue of incomplete data is known to down-bias estimates of total population (Baker et. al., 2012) and to skew estimates of the distribution of population within age/sex or other categories (Baker et. al., 2013, 2014). This paper reports results from a preliminary study attempting to apply principles and tools from Operations Research—specifically the conceptual framework of Lean Principles and the tools of Value-Stream Mapping—to improve geocoding quality. An associated result will be improvement of geospatial micro demographic databases and the population estimates and forecasts that they are used to make (Baker et. al., 2012, 2014).

In this paper, we conceptualize geocoding as a production process not unlike those encountered within the field of Industrial Engineering. As in Figure 2 (next page), local address data that are linked to specific microdata on demographic events including births, deaths, or residential construction are collected. These address strings are standardized to form an initial address list as a candidate pool for geocoding. This preliminary list is presented to a geocoding

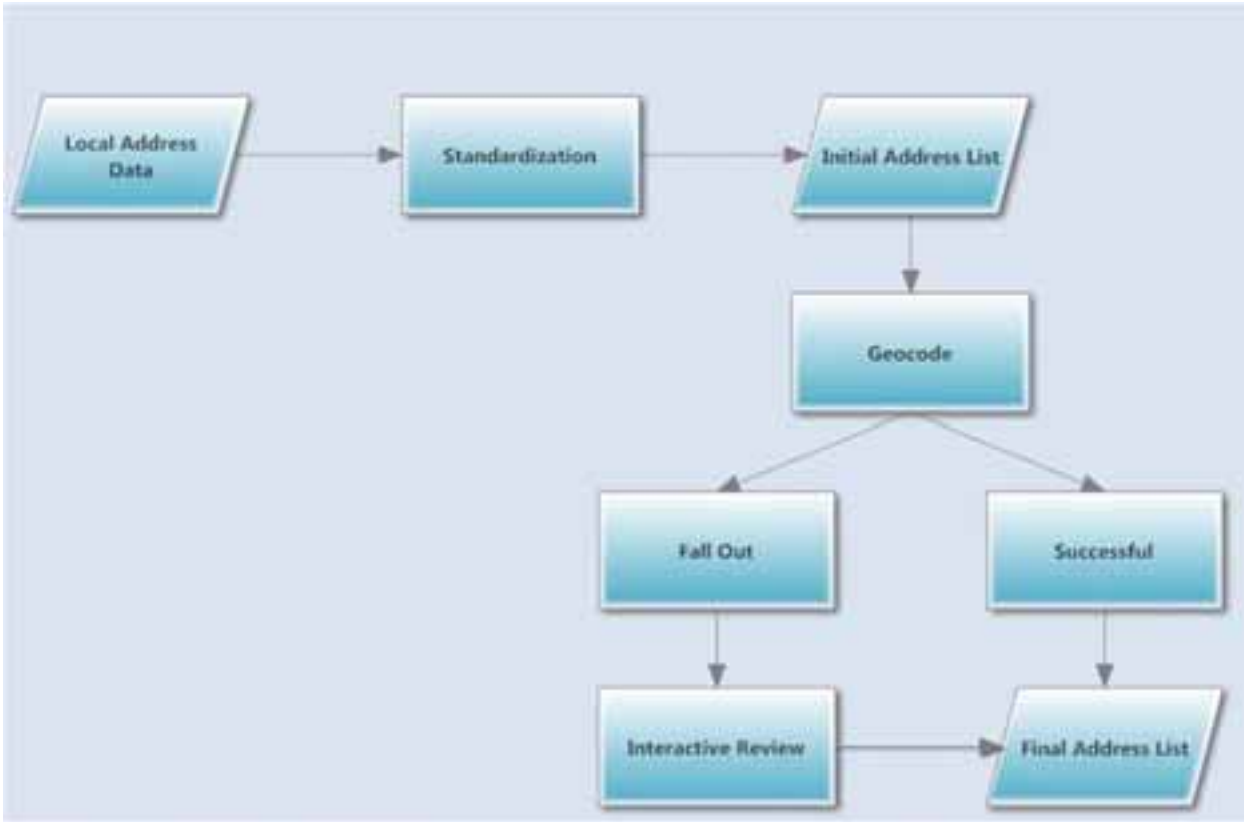

Figure 2. Geocoding as a Production Process

service (in this case that of ESRI's Arc-GIS 10.1) and are either successfully geocoded or not. Those addresses that are not, are grouped as "fall-out" and subjected to an interactive review process that hopefully allows their georeferencing on maps and inclusion in a final address list in combination with those addresses that were originally successfully geocoded. In the engineering world, we could conceptualize this as a "stock and flow" model of the process involved in taking address-based microdata and turning it into georeferenced data that may form the basis of small-area demographic estimates and projections. From such data—we may aggregate results within a specific small-area geography and then proceed to make demographic estimates and forecasts as desired.

The process outlined in Figure 2 may be treated as an engineering problem—with a goal in mind of minimizing the number of records that are found in the "Fall-Out" segment of the figure—or, alternatively, maximizing the number of records included in the "Final Address List". This paper reports our efforts to apply engineering-based perspectives to this problem

using a conceptual framework based on Lean Principles in conjunction with an established tool known as Value Stream Mapping.

MATERIALS AND METHODS

Lean Principles

The 'Toyota Production System' principles were used to reduce errors in geocoding and achieve reduce Fall-out and increase correspondence between the initial micordatasets input into the process and the final address list used to make demographic estimates and projections. The concept of Lean Principles developed by Toyota focuses on reducing waste (*muda* in Japanese)—which detracts from the "value" of a given product in a common-sense fashion. Within the Lean Principles framework a number of important tools are employed to identify and remove waste—value-stream mapping is one such tool that is employed in this study. At the core of the application of Lean Principles reported here is the concept of *Poka-yoke* (error-proofing), which constitutes the main form of waste to be identified in the base process defined in Figure 2. Geocoding errors are treated as a specific form of waste and subjected to error-proofing methods through the application of Value-Stream Mapping.

Value Stream Mapping

Value Stream Mapping includes 4 steps. They are: (1) Construct the Current State Map, (2) Identify Waste, (3) Eliminate the Waste through Error-Proofing (*Poka-Yoke*), and (4) Construct Future State Map. In current state map, the actual process for the production line is described (as in Figure 2—an example of the Current State Map employed in this research). This initial map allows identification of errors as well as measurement of the rate at which such errors accumulate at each step of the production process. It allows potential prioritization of effort (Figure 3)

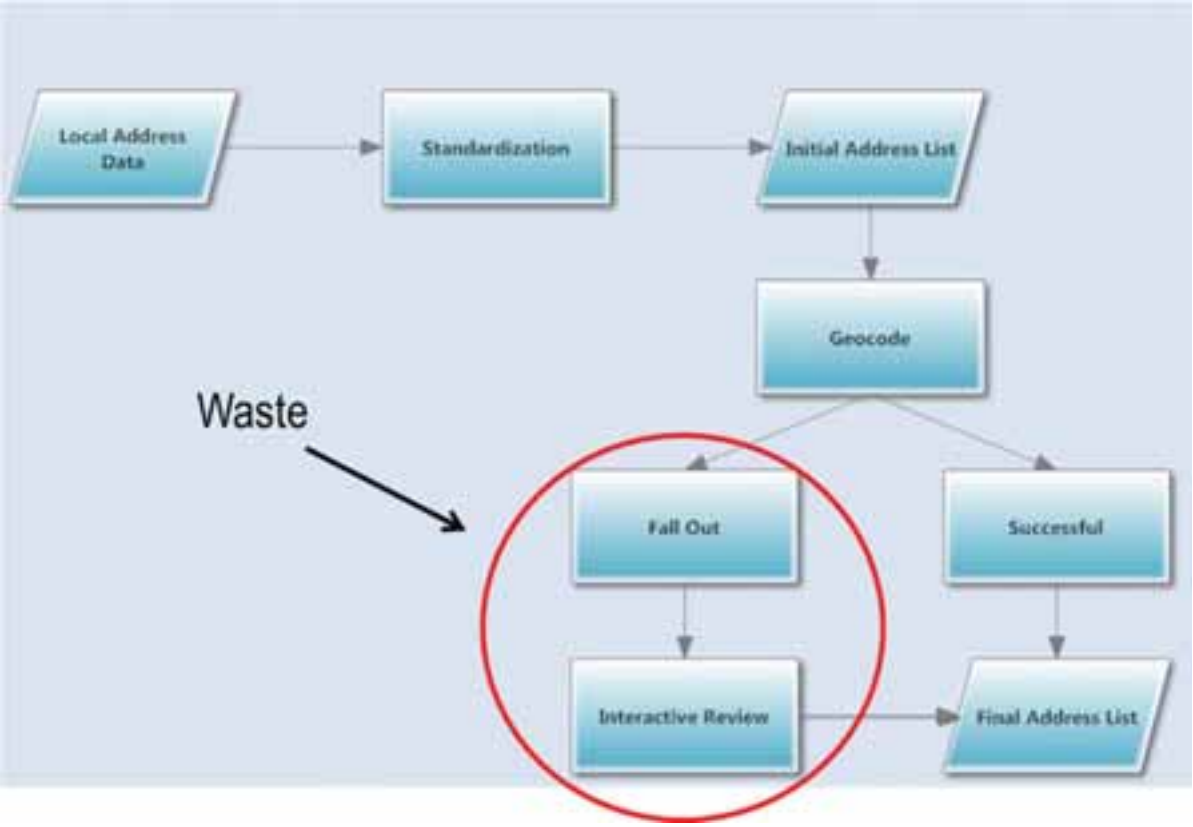

Figure 2: Waste and Prioritization in Light of Current State Map

and identification of the most effective points upon which error-proofing efforts may be focused. It's natural links to cost estimation allow one to estimate the magnitude of the impact of error-proofing efforts on overall waste and link this cost/benefit information to the prioritization of effort.

First Steps: Error Typology and Simulation as Exploratory Tools

Figure 1 shows the current state map of the process used to geocode addresses to support demographic modeling. A first step in this research was to identify the types of errors that prevent successful geocoding of demographic microdata and to build a simulation of the data-generating process based on the Current State Map to facilitate our understanding of the error-generating process. The results of this initial phase would then be used to target error-proofing

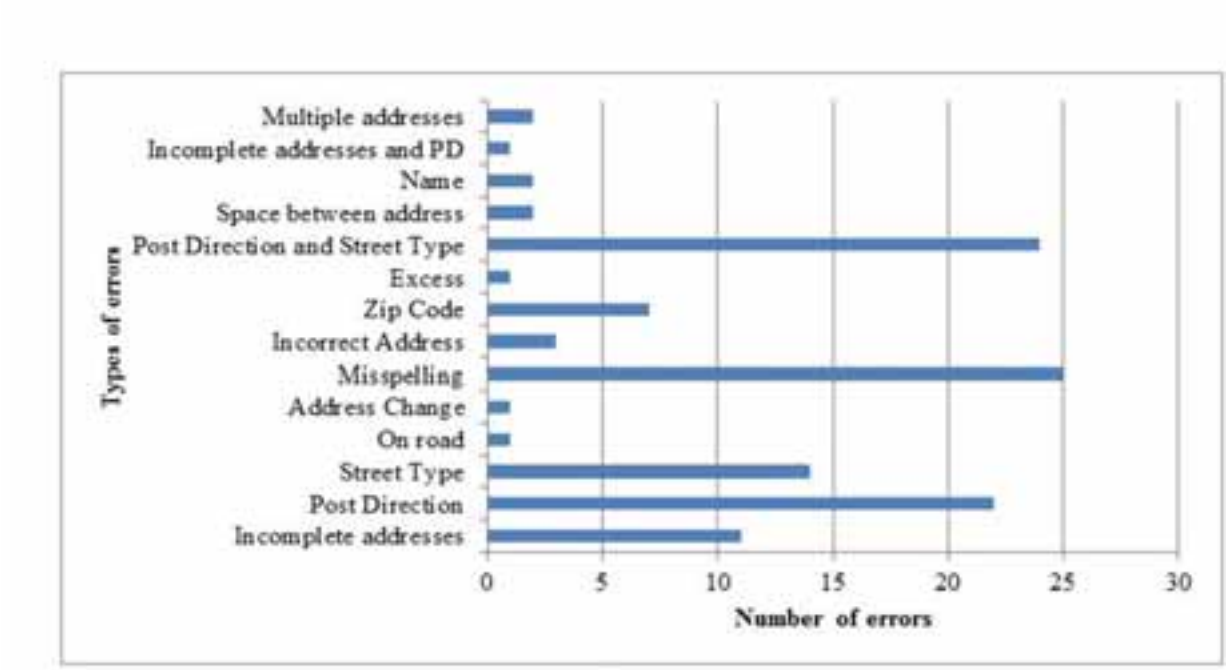

Figure 3. Number and Frequency of Types of Geocoding Errors, Sandoval County (2011)

algorithmns in the most efficient manner possible. Figure 3 reports the results obtained from an initial case-study for Sandoval County, NM for the year 2011. It reports results for addresses generated from a number of different administrative data sources (building permits, birth and death records, driver's licenses, etc.) for the county as a whole. While incomplete addresses (deletions of sections of the string) present a significant source of error, misspellings and errors in street type or post-directional designation appear to be the largest sources of error in the case-study: these errors are responsible for the greatest number of records that ended up in the "Fall Out" bucket of the Current State Map provided in Figure 1.

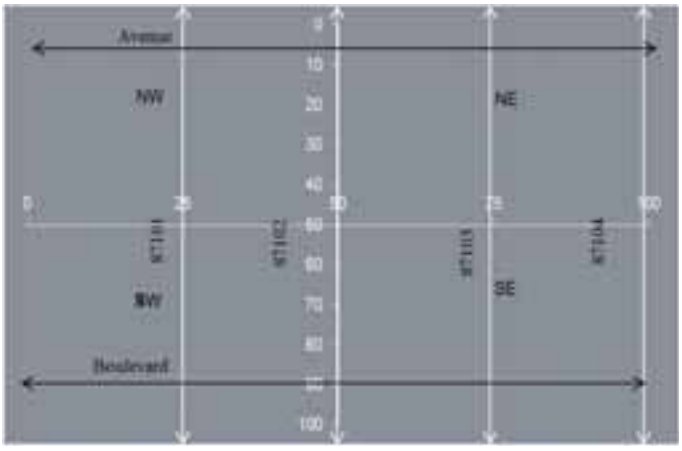

Figure 4. Anytown USA Simulation

Further insights were gained through simulation (Anytown USA—Figure 4) designed to mimic the geocoding process and comprised of four zip-code quadrants and eight intersecting streets—four running in each direction from North to South and East to West. Addresses were numbered 1 to 100 along each of these streets and post-directionals

were assigned based on the four natural quadrants of the City. This provided a list of pure addresses that were 100 percent accurate and error free. This address list was then systematically corrupted according to the errors observed in the Sandoval County case study to arrive at set of 1000 addresses with a specific fraction corrupted in a way that was clearly understood. The fairly obvious conclusion of this process was that if a gold-standard address reference list—or address locator built from such—existed, then the correction of misspellings and incorrect street types to this reference list could be accomplished as the most efficient means of minimizing geocoding fall-out or, inversely, maximizing geocoding success rates.

We suspect that this appears to be a trivial result to those experienced with the process of geocoding—but one clarity reached as a result of the simulation process was that correcting discrepancies between an address list to be geocoded and the reference list of road networks against which we would attempt to geocode them should be a fruitful avenue upon which to focus efforts. In our dataset, addresses are typically presented against address locators that are built from road networks that include information on the range of addresses available within a specific road segment, the name and street type of the road segment, and its pre and post-directional designations. We reasoned that correcting our ungeocoded addresses to this reference list of street names and types we could enhance our ability to geocode individual records. While this tactic would not solve all of our problems—for example, it would not fix incorrect or incomplete house numbers, defective or incomplete zip-code designations, or the incompleteness of the actual electronic road networks that serve as the basis for georeferencing—it would address the two biggest types of errors identified in our case-study and likely solve much of our problem with completeness. As a last step in the simulation, we utilized a well-known string comparator and correction-to-reference procedure in the SAS software package (PROC SCAN) as a means to correct the simulated defective addresses to the original list and achieved a predictabely high (nearly 100 percent) success rate. While apparently trivial, the computer simulation was an important tool for gaining clarity about how to proceed with correction efforts.

A Trial Solution: Restandardization as a Tool for Correction

Based on the results of the case study and simulation, we proposed tto attempt standardization of the ungeocoded address to the appropriate road network as a form of

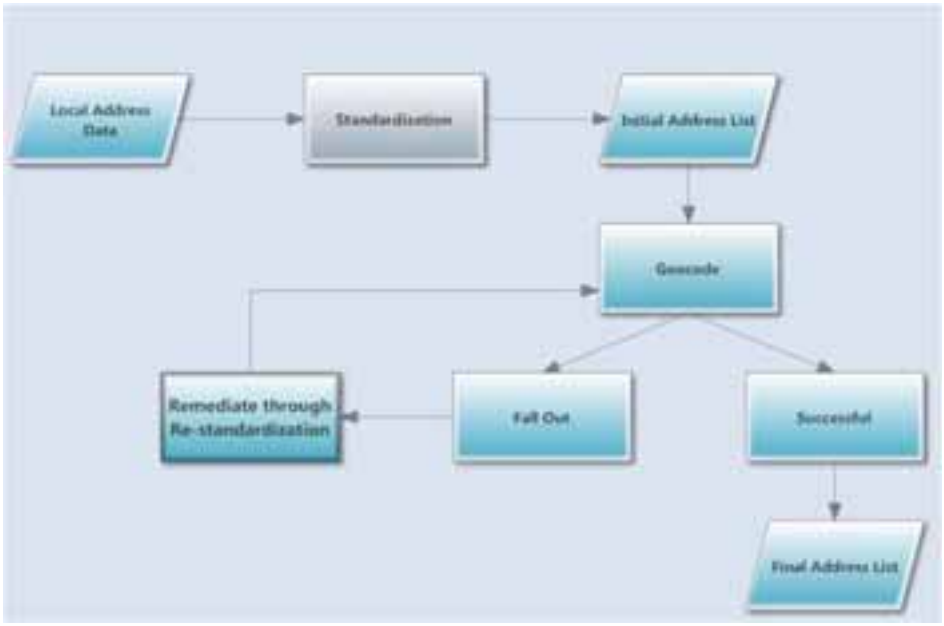

Figure 5. Restandardization as Error-Proofing

error-proofing (Figure 5). Addresses that fell out of the initial geocoding were then subjected to standardization to the road-network-derived candidate address lists then re-presented for geocoding. The process of restandardization is presented in Figure 6. This process would not correct for incomplete addresses, zip-code, or house number defects; however, it was anticipated that it would dramatically improve overall geocoding success rates. We set a target of improving geocoding match rates from the 78% percent observed in the original geocoding to 90%.

As in Figure 6, the restandardization process focused on street names and post-directionals (pre-directionals are rare within Sandoval County)—and therefore, did not remediate defects or omissions in house number of zip-code. The possible address universe, to which the addresses that were not geocoded were

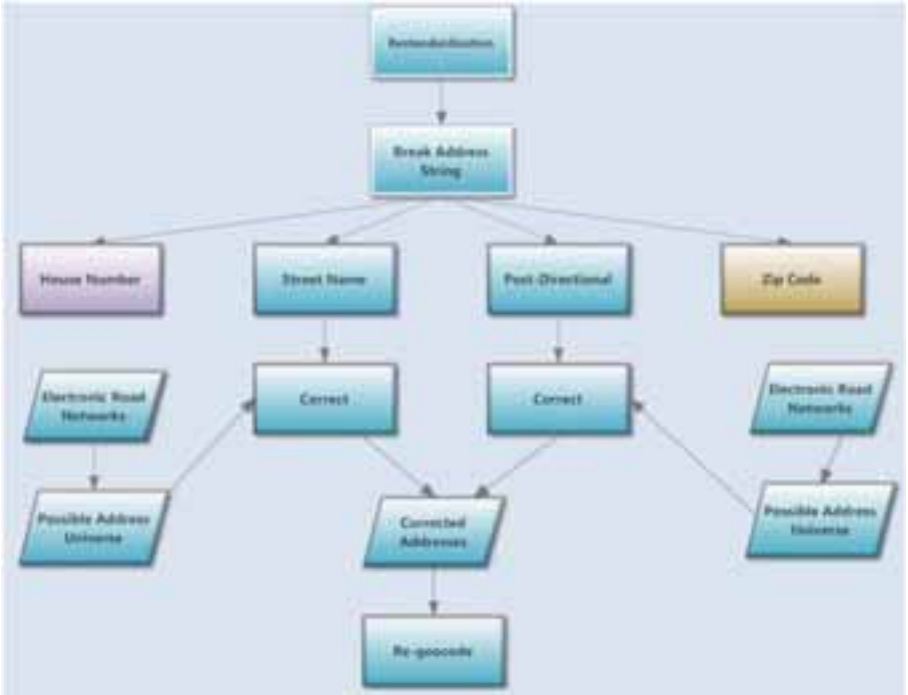

Figure 6. The Process of Restandardization

standardized, were derived as previously-described from the electronic road networks that were used to form address locators.

Results: Geocoding Improvements and Associated Cost-Offsets

After the restandardization process was completed, 79.00 percent of the previously ungeocoded addresses that were presented to the ESRI 10.1 geocoder were successfully mapped. This moved the overall geocoding success rate of the entire process to 95.00 percent. While 95.79 of the ungeocoded addresses were successfully remediated to correct street names and post-directionals, one-hundred percent geocoding at the street level was not obtained because of defective or omitted house numbers and incomplete road networks. Anecdotally—in many cases addresses could be located on Google Earth—suggesting that incomplete road networks remain an important source of "Fall-Out".

In the final tally, geocoding success was improved by 17 percentage points—from 78.00 percent to 95.00 percent; this will have an important impact on both the accuracy of the population estimates and forecasts produced using these data as well the cost of formulating it. While a missed birth may lead to an underestimate of one person and a missed death to an overestimate of one person, missed geocodes of housing units that are often used to estimate migration for small area demographic estimates and forecasts (Baker et. al., 2012, 2013, 2014) have a multiplicative effect. A single housing unit can lead to a mis-estimate of as many as three persons (Baker et. al., 2012, 2014). To get a sense of the magnitude of such effects, one need only look at residential permanency. Across most studies, upwards of 80.00 percent of the population lives in the same residence as the previous year—this suggests a flux of 20.00 percent of persons that may be shifting from one region to another—or even from one census tract to another. By far—we should expect estimates of net-migration to be the most sensitive to geocoding errors and with as many as twenty percent of persons moving within a given year, the impacts of geocoding errors on demographic estimates and forecasts might be very large.

It is clear that the remediation process reported here will improve the accuracy of demographic estimates and forecasts made using geocoded data. The potential cost-offset is also extremely impressive. Table 1 (following page) breaks down the cost savings associated with

Activity	Number of Records	Cost Per Record	Total Cost	Cumulative Cost	Percent of Data Capture
Geo-coding	20,379	$0.50	$10,189.50	$10,189.50	78.00
Interactive Hand-Matching	2,874	$1.00	$2,874.00	$13,063.50	88.00
Aerial Imagery Review/Google Maps	1,725	$5.00	$8,625.00	$21,688.50	94.65
Fieldwork	1,149	$15.00	$17,235.00	**$38,923.50**	100.00
Geo-coding	20,379	$0.50	$10,189.50	$10,189.50	78.00
Proposed Process	4,540	$.23	$1,023.75	$11,213.25	95.28
Left Imagery/Fieldwork	1,208	$10.00	$12,080.00	**$23,293.25**	100.00

implementing the process for the collection of 20,379 addresses geocoded as part of this study within Sandoval County, NM. The table provides counts of records captured in each step in the process, the cost per record of each type of activity, the total cost of the activity, and the cumulative cost and percent of data capture. These correspond to the processes described in the Current and step 1 Future State Maps.

The total cost offset is estimated to be over $15,000.00—out of a total cost to the original process of over $38,000.00. The cost offset is approximately 40.00 percent. While in this comparison seventeen percentage points were gained in terms of coverage—in practice the post-hoc hand-matching, imagery analysis and fieldwork-based remediation would likely have narrowed this difference significantly, *but at a substantially increased cost.* Even if equivalent coverage were obtained in the end, a forty percent cost savings is an important and welcome result.

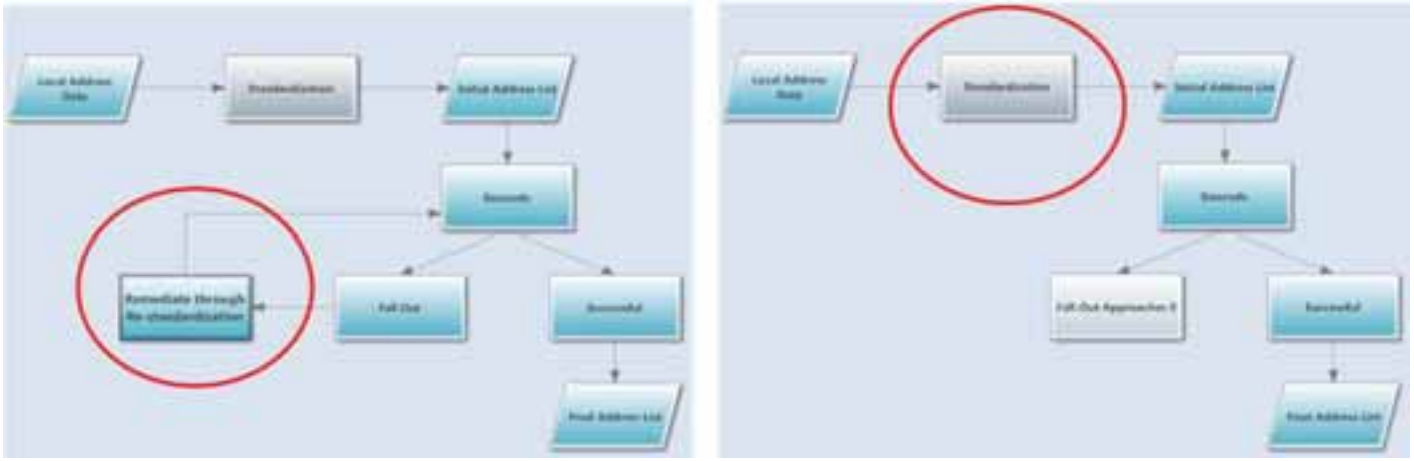

Figure 7. Further Improvements: A Future State Map for Moving Forward

Conclusions

The results of this research certainly encourage further implementation of methods from industrial engineering/operations research to the process of geodatabase development. Figure 8 illustrates the Future State Map suggested by this research. In the first proposed Future State Map, we conceptualized the method as aiming at dealing with Fall-Out in a two-step process. The final Future State Map (right side of Figure 7) suggests that this iteration is pointless—instead, standardization of input address lists to that suggested by electronic road networks used to form address locators may happen as an initial step. There is no value added in postponing this step to deal with Fall-out as it is extremely inexpensive to add this step of routines into existing standardization processes that we already implement at the front of the Current State Map. As this constitues only a small set of additional loops in SAS, the delay in time of programming costs is negligible. Standardizing up front should have no effects on accuracy other than those observed in the two step process. The up-front cost of parsing and exploring data to optimize programming routines on a county-by-county basis introduces an up-front research cost that will easily pay for itself beyond its first year of implementation if the results of this study hold. The 20,379 addresses geocoded in this research exercise are only a fraction of the number of geocodes processed by our group in a year (which can reach as many as 500,000 to 600,000 records). Clearly, the algorithmic approach advocated in the Future State Map of Figure 7 is a hugely cost-effective effort.

The process may be capable of further improvements as well. In the current study, no attempts were made to deal with pre-directionals, defective or missing house numbers, defective or missing zip codes, or post-office boxes. In each case, additional progress is anticipated in future research applying Lean Principles and Value-Stream Mapping to geodatabase development. In future efforts, we plan to pilot methods of house number imputation to deal with missing or incomplete house numbers, machine learning algorithms to deal with updates in the range of addresses associated with a specific street segment and to deal with "out-of-range" or missing house numbers. We plan to employ linkage algorithms to attempt to deal with how to geocode PO Box addresses into the correct census tract or block, rather than zip-code (which may not even reflect residence zip-code in many cases). Efforts will also be focused on dealing with errors in assignment of zip-code, temporal changes in zip-code status, and missing zip-code elements.

While promising, the results encountered in the current study should be extended to other geographic areas—most notably those that are characterized by different challenges and error structures than those associated with Sandoval County. Most new addresses in Sandoval County are urbanized, located within the fastest-growing city in the State: Rio Rancho. This means that incomplete road networks, road-name changes, and simple misspellings should predominate explanations of incomplete geocoding (Fall-Out). In other examples, the impact of algorithmic corrections may be much less pronounced since incomplete road networks and PO Boxes likely present a much greater challene to data capture through geocoding. In urbanized areas, substantial process can be made through algoritmic means. In less urbanized areas, it remains to be seen if such a process may have such large-scale impacts. No algorithm based on the address side can fill in missing road networks—in these areas, it is likely that digitizing roads and investing in significant field work may remain as the only true options for approaching 100 percent data capture. The gains in accuracy of such efforts may not be that large in terms of improving demographic model accuracy (Baker et. al., 2012, 2013, 2014; Smith et. al., 2001; Swanson and Tayman, 2012), suggesting that they may not be worth the additional effort in this specific context. Only future research will resolve such issues. The current research suggests that it is worth the effort to figure out how much improvement may be made across a large variety of settings.

References

Baker J, Ruan XM, Alcantara A, Jones T, Watkins K, and additional authors (2008) Density-dependence in Urban Housing Unit Growth: An Evaluation of the Pearl-Reed Model for Predicting Housing Unit Stock at the Census Tract Level. Journal of Economic and Social Measurement, 33,155-163.

Baker J, Alcantara A, Ruan XM (2011) A Stochastic Version of the Brass PF Ratio Adjustment of Age-Specific Fertility Schedules. PLoS One. 6(8):e23222. doi:10.1371/journal pone0023222

Baker J, Alcantara A, Ruan XM, Watkins K (2012) The Impact of Incomplete Geocoding on Small Area Population Estimates. Journal of Population Research, 29, 91-112.

Baker J, Alcantara A, Ruan XM, Watkins K, Vasan S (2013) A Comparative Evaluation of Error and Bias in Census Tract-Level Age/Sex-Specific Population Estimates: Component I (Net-migration) vs Component III (Hamilton-Perry). Population Research and Policy Review. 32:919-942.

Baker J, Alcantara A, Ruan XM, Ruiz D, Crouse N (2014) Sub-County Population Estiamtes Using Administrative Records: A Municipal-Level Case Study in New Mexico. In: Emerging Techniques in Applied Demography. Nazrul Hoque and Lloyd Potter, editors. New York: Springer.