

Geocoding Truck Stops Update

William Co

2025-07-27

This report documents the geocoding validation process, highlighting discrepancies arising from integrating data sources like Yelp and Yellow Pages. Despite matching entries by phone number, these platforms introduce a high rate of false positives. We compare point discrepancies with and without these sources and visualize the magnitude of spatial discrepancies. The report discusses the potential causes of these inconsistencies, ranging from business relocations to outright geocoding errors, and discusses strategies for post-processing, including centroid averaging and external geocoding API cross-checks.

Table of contents

| | | |
|----------|---|----------|
| 1 | Introduction | 1 |
| 2 | Discrepancy Analysis | 2 |
| 2.1 | Without Yelp and Yellow Pages | 2 |
| 2.2 | With Yelp and Yellow Pages | 2 |
| 2.3 | General Overview | 2 |
| 3 | Next Steps | 2 |

1 Introduction

During the process of geocoding truck stops, we observed a significant number of false positives. Although Yelp and Yellow Pages entries were matched by phone number, this method still introduced a substantial number of false positive matches.

The complete dataset can be accessed [here](#), which includes the latitude and longitude values used for geocoding.

2 Discrepancy Analysis

2.1 Without Yelp and Yellow Pages

If we remove the all Yelp and Yellow Pages entries, we observe point discrepancies illustrated in [this HTML map](#). This file displays only those locations where the geocoded points differ by more than 1 km.

2.2 With Yelp and Yellow Pages

When we include the Yelp and Yellow Pages entries, the discrepancies increase significantly, as shown in [this HTML map](#). The discrepancies are so numerous that the HTML visualization becomes difficult to interpret due to the overwhelming number of false matches. To take this into account, the current visualization has been redone to display only point discrepancies with distances greater than 16 km.

2.3 General Overview

For a high-level overview of the geocoded points, see [this map](#), which provides a visualization of all the coordinate data regardless of accuracy.

3 Next Steps

This issue could potentially be addressed through **manual review**, although this approach has limitations. Some discrepancies are not clearly identifiable as errors (“splitting hairs”), which complicates the process. At this stage, an **important research judgment** must be made.

The observed differences might simply reflect the **age of the locations**. Perhaps some businesses have moved. Alternatively, the discrepancies could stem from **genuine geocoding errors**.

One possible approach is to compute and use the **centroid** of the available coordinates. Another option is to introduce a **third validation step**. For example, we could use address observations to cross-check the coordinates using **geocoding API** such as Google Maps and I-Exit coordinate data as well. I-Exit validation would be straightforward but using google maps geocoding API, would also run into similar concerns about age discrepancies of locations.

Another potential step is to perform string matching between all phone number based matches against the original data. However, this approach would likely require research judgment, as string matching is not perfect as well.

While we have outlined several potential next steps in this paper, the decision on how to proceed ultimately involves a degree of research judgment, more an interpretive choice than an objective one. As such, any input or guidance would be greatly appreciated.