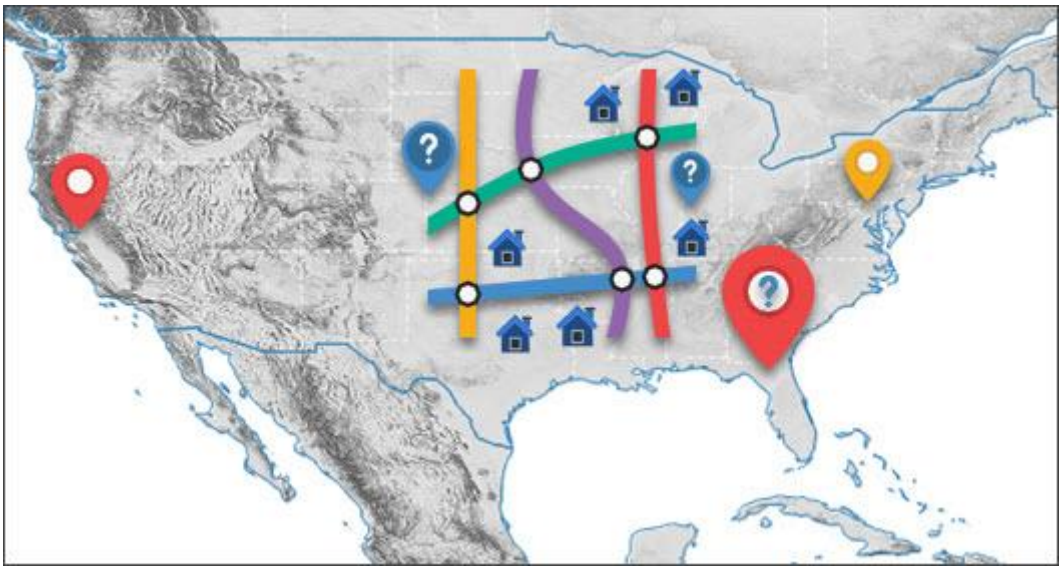
Geocoding DoubleCheck: A Unique Location Accuracy Assessment Tool for Parcel-level Geocoding

Geocoding is a process of converting an address (typically in the form of street number, street name, city, state and country) into a physical location on the globe (often expressed as latitude and longitude). It is one of the most popular spatial analysis techniques, and almost all major GIS software (e.g. QGIS, ArcGIS and MapInfo) and web mapping portals (e.g. Google Maps, Bing Maps and Apple Maps) provide geocoding tools and API options. But the big question is how reliable is for the geocoding result from each geocoder.

Geocoding (location) accuracy is fundamental to the exploration of **location-sensitive** or **location-centric business intelligence**, and important for applications such as asset management and risk analysis. Currently very few assessment tools on this exist. Without the confidence on the location accuracy, subsequent analytics and results would be seriously compromised.

We have developed **Geocoding DoubleCheck**, a unique tool for assessing the location accuracy of the widely used parcel-level geocoding. The current coverage is for 48 contiguous U.S. States.



Parcel-level Geocoding vs. Building-level Geocoding

In a basic form, geocoding is realised by interpolating the address range numbers specific to individual street segments (i.e. **street-level geocoding**). For advanced geocoding, the following two levels can be categorised (Figure 1):

Parcel-level geocoding: The geometric centroid of a land parcel is used to represent the location of an address. This is a popular and increasingly-applied form for geocoding in some developed countries, e.g. about a half dozen private companies offering land parcel data with varying completeness in the U.S., and the open release of the Geocoded National Address File in Australia.

Building-level geocoding: The geocoded location is indicated by the building footprint of a physical building and/or a point within (also known as **rooftop-level geocoding**). The latest pursuits are certainly geared towards this (e.g. the development of building footprint data by CoreLogic in the U.S. and by PSMA in Australia), but it needs significant investment for a very large geographic coverage. Google produces and holds very comprehensive data sets on digitised building footprints, as seen in Google Maps, but they are proprietary and not open to external analysis.

Obviously, for many location-centric applications such as site-level risk analysis, the location accuracy of parcel-level geocoding is still not enough and what is required is the geocoding at the physical building level.

Figure 1: Comparison between parcel-level geocoding (left) and building-level geocoding (right). Location: Tompkins County, NY. (Acknowledgements: In this document, the sample land parcel and building footprint data are obtained from the Tompkins County GIS Division; high-resolution aerial imagery is from the USDA NAIP series.)

The difference between the two advanced geocoding levels above can be more easily appreciated if **high-resolution aerial or satellite imagery** is superimposed (Figure 2). Imagery provides a rich context and clearly shows various land covers. In this demonstration, the limitation of the parcel-level geocoding for location-centric applications such as risk mapping is obvious. For example, a house might be far away from adjacent fire-prone forest but its parcel centroid could be within forest. (Figure 3 shows another example in risk mapping: a house might be located on a high ground 50m away from a nearby river but its parcel centroid could be in a very flood-prone zone.)

Figure 2: Land parcels and centroids (polygons and dots in orange) and building footprints (white polygons) are overlaid with 1m-resolution aerial imagery. The spatial discrepancy between parcel-level geocoding and building-level geocoding is shown.

The Use of High-Resolution Imagery and Classified Vegetation

High-resolution imagery and classified vegetation are essential for the development of **Geocoding DoubleCheck** in assessing geocoding location accuracy (Figure 3). Ideally, high-resolution land cover maps can be readily sourced but they usually do not exist for a very large territory at this time. Figure 4 compares classified vegetation at 1m resolution and classified land covers from the 2011 National Land Cover Database (NLCD) at 30m resolution, and it is clear that for this type of application, **the use of high-resolution imagery and classified land covers is a prerequisite**.

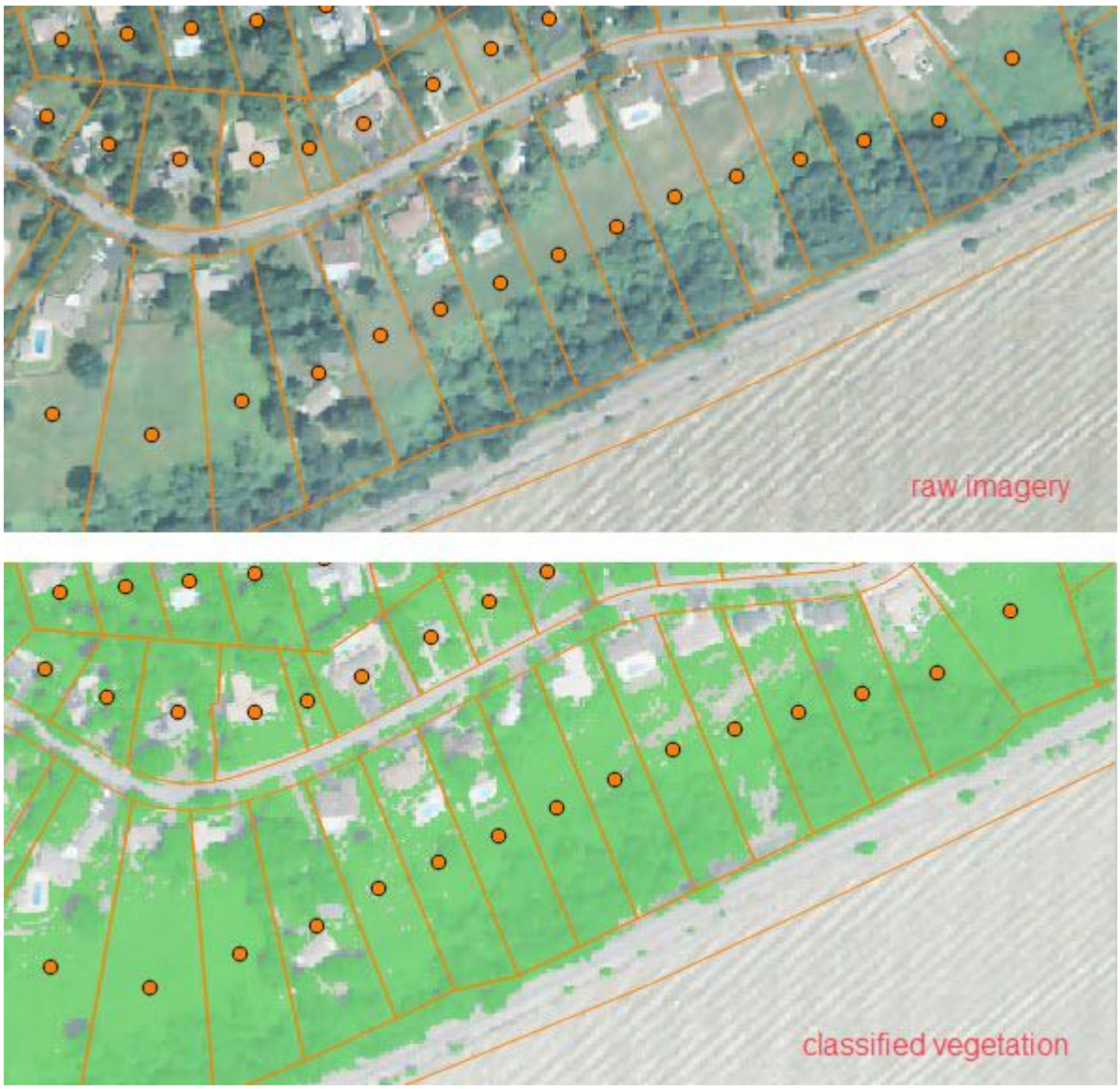


Figure 3: Parcel-level geocoding result superimposed with raw imagery (top) and classified vegetation (in green colour, bottom). The classified vegetation map can be used to determine if the land cover for the geocoded spot (parcel centroid) is vegetation or not.

Figure 4: Comparison between classified 1m-resolution vegetation (left, in green) and the classified 30m-resolution land covers from NLCD 2011 (right, colours indicating different classified land covers). Land parcels (white polygons) and their geometric centroids (black dots) are superimposed.

Geocoding DoubleCheck takes advantage of the new, comprehensive high-resolution (1m) digital aerial imagery for 48 continuous U.S. States (from the USDA NAIP 2015/2014/2013 series) and the recently classified unique vegetation dataset by BigData Earth (link). We are now able to determine if a geocoded location at the parcel centroid is vegetation or not. If it is vegetation, the geocoded spot is subject to further location scrutiny or improvement. As the imagery we analysed is from agricultural growing seasons or with "leaf on" conditions, the presence of vegetation is a universal feature and a good indicator for open space. Although not perfect accuracy, like any other image classification tasks, it does serve as a highly relevant and reliable basemap that can be used to determine site land cover conditions for the overwhelming majority of geocoded locations.

Application Example: Advancing Loss Modelling and Risk Mapping

We identify and offer three application levels in relation to the Geocoding DoubleCheck product:

Level 1: Summary

If only a very small number of geocoded addresses (for various asset and exposure types) are involved, a user may rely on high-resolution imagery to cross check their true locations and make on-screen adjustment manually.

Geocoding DoubleCheck – Summary software tool is ideally used to analyse 1,000s or even millions of geocoded addresses in an automated workflow and report the % of addresses that might just hit open space, i.e. vegetation. A typical application for this is **catastrophe loss modelling** used by the (re)insurance industry, where a large number of geocoded addresses (for exposure or portfolio data) are routinely analysed. Location accuracy is very important as a few meters could mean a different hazard level. Better location accuracy for exposure data has long been demanded by the industry, and we are now able to automatically assess the location accuracy of a large quantity of exposure data used in aggregate loss modelling. **The use of geospatial big data analytics and high-resolution digital imagery has now made this large-scale implementation possible**.

It has been widely recognised by academics and risk management practitioners that exposure location accuracy is a major source of uncertainties in loss modelling. Imagine that every exposure input is filtered through software like the **Geocoding DoubleCheck – Summary** tool, the end user would be much more confident on the exact input that has greater location accuracy. This will certainly reduce uncertainties and make catastrophe loss modelling more transparent and powerful.

Software inputs and outputs:

- Typical input for the software tool is very simple and requires four basic fields: unique ID for each exposure record, geocoded latitude, geocoded longitude, and State name.
- Typical output is a summary table showing the % of geocoded addresses that hit vegetated areas, and the % of geocoded addresses that have greater location accuracy and been analysed for loss calculation. This can be reported by regions (e.g. states) or by any other input attributes.
- We offer customised processing services and/or standalone software tools with terabytes of classified vegetation basemaps stored on external hard drives. Web services are not available at this time.

Level 2 – Refinement

As illustrated in Figure 3, if the distribution of classified vegetation is known, the centroid of each land parcel can be adjusted relative to non-vegetation areas and a new geocoded location in close proximity to a physical building may be determined. Indeed, this type of analysis can be extended by considering any other non-building land covers (e.g. open water and road). We offer **advanced image processing services** to facilitate such location adjustments and refinement **at the land parcel level**. In this case, in addition to high-resolution imagery, land parcel data need to be outsourced as they provide predefined control boundaries for detailed land cover classification.

Level 3 – Supplement

This examines geocoded locations by considering many terrain and environmental metrics at a site level. It provides an opportunity to double check and diagnose the surroundings of each individual site from multiple perspectives.

A typical example is **site-level risk mapping**. This application overlaps with another data product we are developing (i.e. New Generation Property Location Information, link), which measures and quantifies a range of terrain and environmental attributes, including land covers, terrain, climate change and natural hazards.