Capture and Visualization of Change In U.S. Census TIGER/Line Data: A Business Intelligence Approach

Northeast Arc Users Group (NEARC)

May 16, 2022

Brian Hebert, Data Architect

ScribeKey, LLC

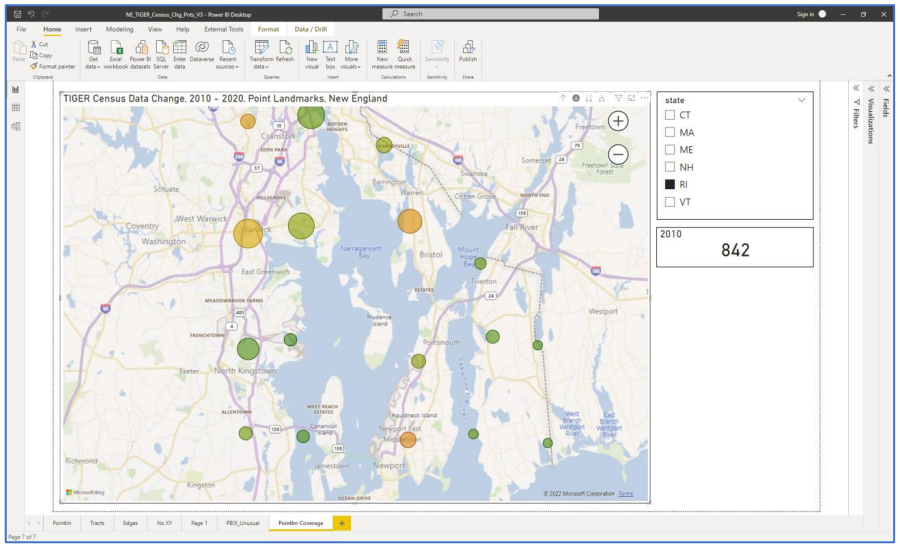
Overview

ScribeKey has developed a technique to capture and present change in high volume geospatial vector data.

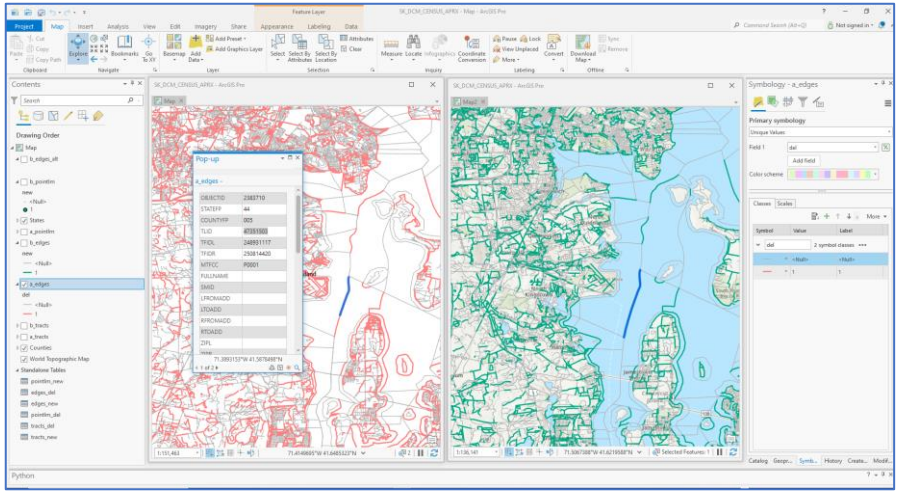
Results can be viewed using both MS Power BI and GIS platforms (ArcGIS, QGIS), in an ad-hoc or systematic (map book atlas) manner.

In this presentation, the capture and presentation of data change metrics is described, using a subset of 3 layers from TIGER/Line data: Point Landmarks, Edges, and Census Tracts, and studying the change that occurred between the 2010 and 2020 datasets.

These tools and techniques have been developed and refined over the past 10 years, working primarily with HERE Technologies (formerly Navteq) datasets.



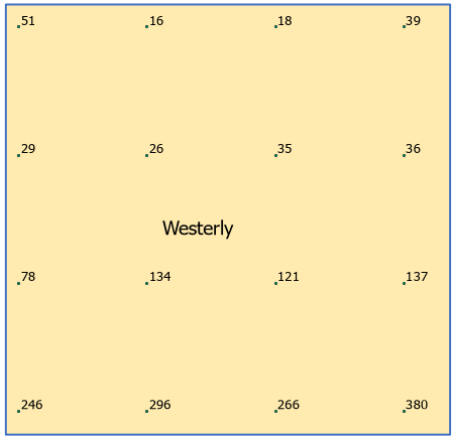
MS Power BI: Hierarchical Change Points



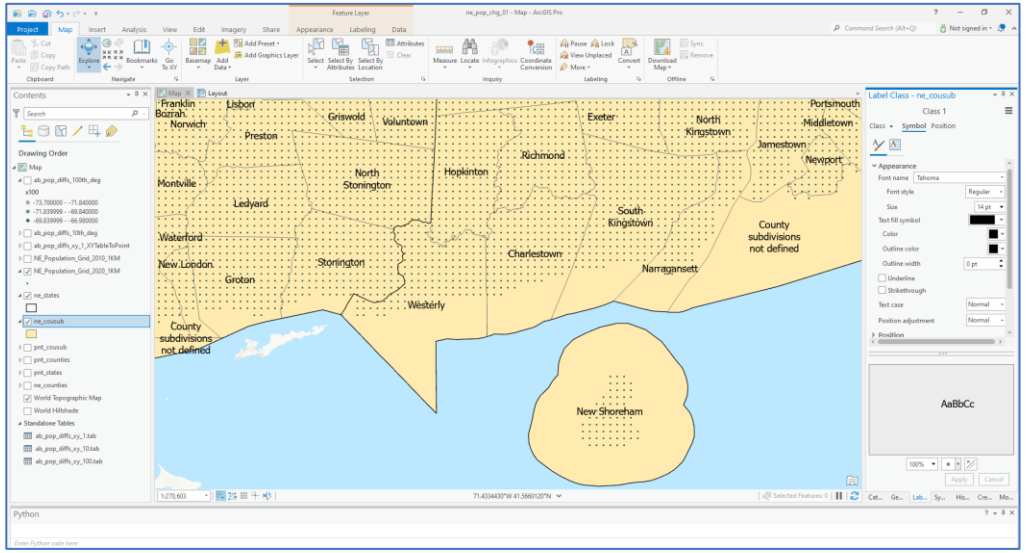
ArcGIS Pro: Feature Detail

Point Grid Maps: Population

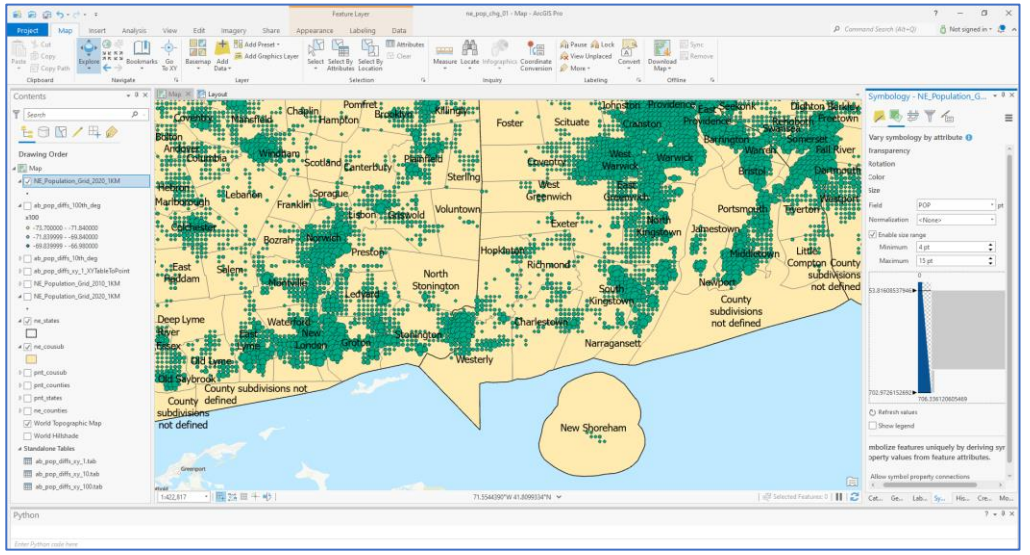
- ScribeKey's main cartographic data product used for capturing and viewing generalized data change metrics is the Point Grid Map.
- Here, to familiarize attendees with this basic concept, an example of a Point Grid Map is viewed in ESRI's ArcGIS Pro.
- This data comes from the Gridded Population of the World. Each point, at a 1-kilometer resolution, has a numeric value indicating population. The numeric value can be though of as a result of the nearest population features (people) being 'snapped' to the grid point.
- Point Grid Maps provide a simplified, generalized view of the underlying data detail, and allow viewers to locate concentrations and clusters.
- Cartographic techniques for highlighting the main aspects of Point Grid Maps include filtering, sizing, and labeling.



POP Value as Label

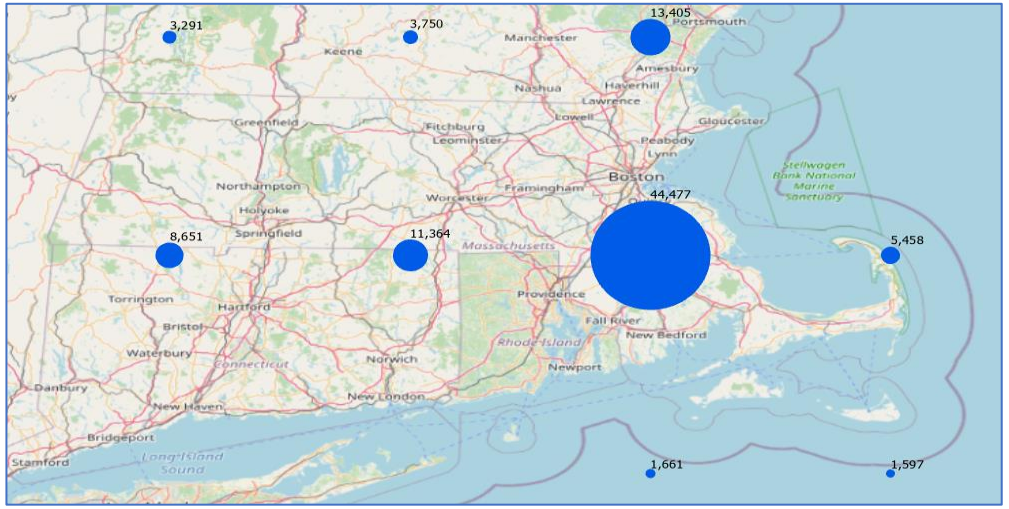


CT/RI Coast – All Points, 2020

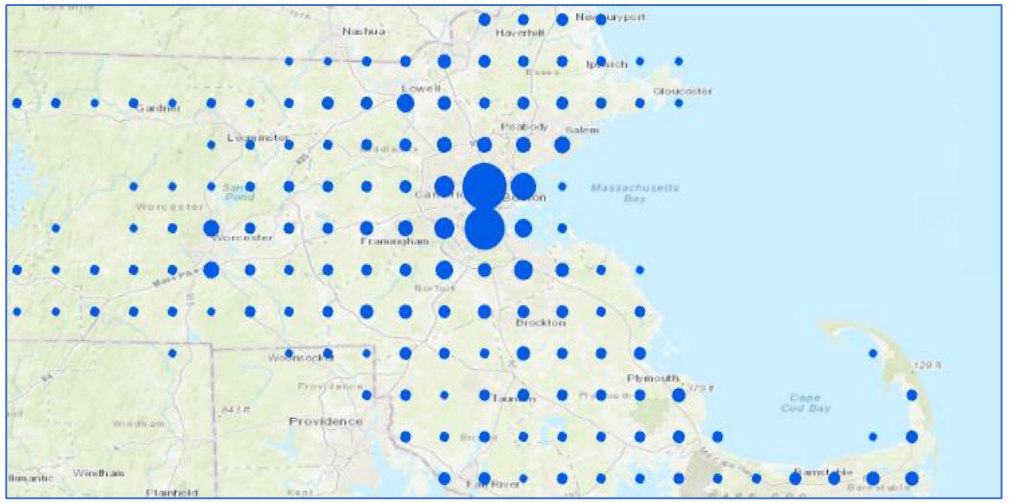


Filtered >= 10 Features

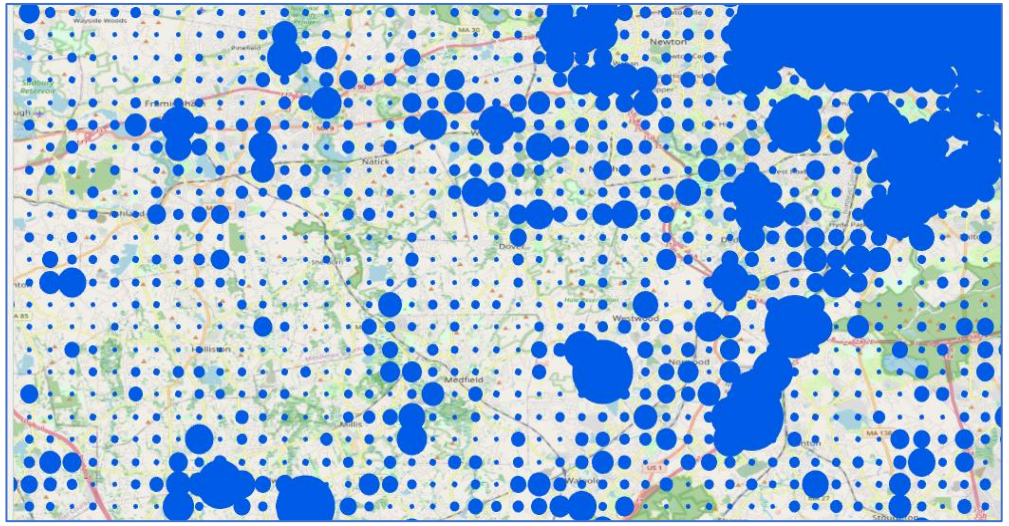
Point Grid Maps: Geospatial-Features/Hierarchy Sizing and Filtering



1 Degree



1/10 Degree



1/100 Degree

ScribeKey applies this same Point Grid Map technique in capturing geospatial feature distributions, and at multiple levels in a 3-tier hierarchy using 1 degree, 1/10th degree, and 1/100th degree resolutions.

Here the feature counts of TIGER/Line Edges polyline features have been 'snapped' to the nearest point.

Again, the Grid Map Points have been filtered and sized to help highlight feature concentration and distribution.

Pivot Tables

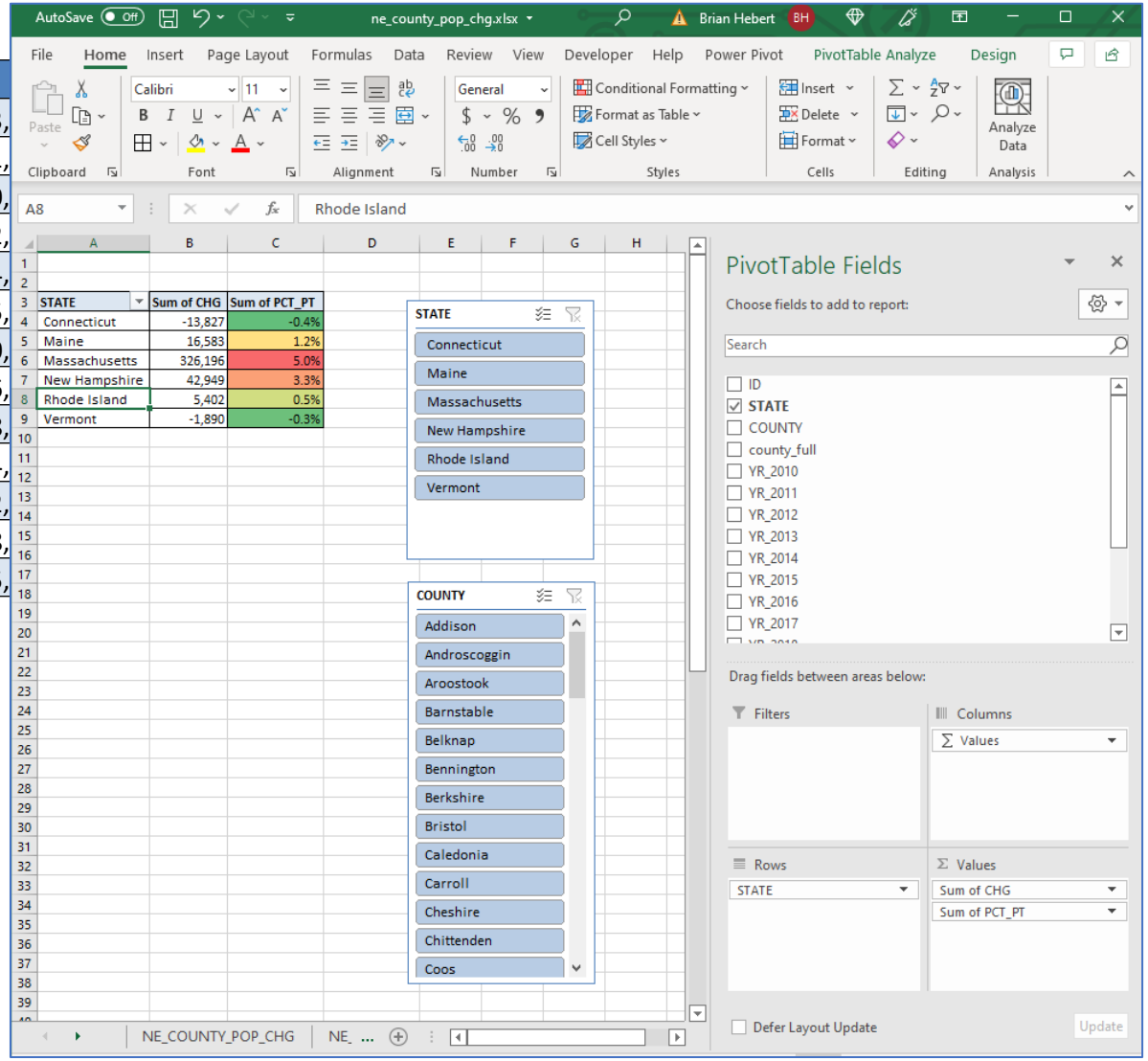
ID	STATE	COUNTY	COUNTY_FULL	YR_2010	YR_2019
1	Connecticut	Fairfield	Fairfield County	919,355	943,
2	Connecticut	Hartford	Hartford County	895,236	891,
3	Connecticut	Litchfield	Litchfield County	189,763	180,
4	Connecticut	Middlesex	Middlesex County	165,616	162,
5	Connecticut	New Haven	New Haven County	863,357	854,
6	Connecticut	New London	New London County	274,004	265,
7	Connecticut	Tolland	Tolland County	153,239	150,
8	Connecticut	Windham	Windham County	118,544	116,
49	Rhode Island	Bristol	Bristol County	49,818	48,
50	Rhode Island	Kent	Kent County	166,030	164,
51	Rhode Island	Newport	Newport County	83,176	82,
52	Rhode Island	Providence	Providence County	627,838	638,
53	Rhode Island	Washington	Washington County	127,097	125,

Base Data Table

The use of ScribeKey's multi-level point grid hierarchy for viewing feature distribution detail on maps is akin to the use of a Pivot Table (here in MS Excel), which constitutes a fundamental tool in Business Intelligence (BI).

Increasing detail can be seen by 'drilling down' into finer grained aggregations. Higher level summaries can be seen by 'rolling up' through the aggregation levels.

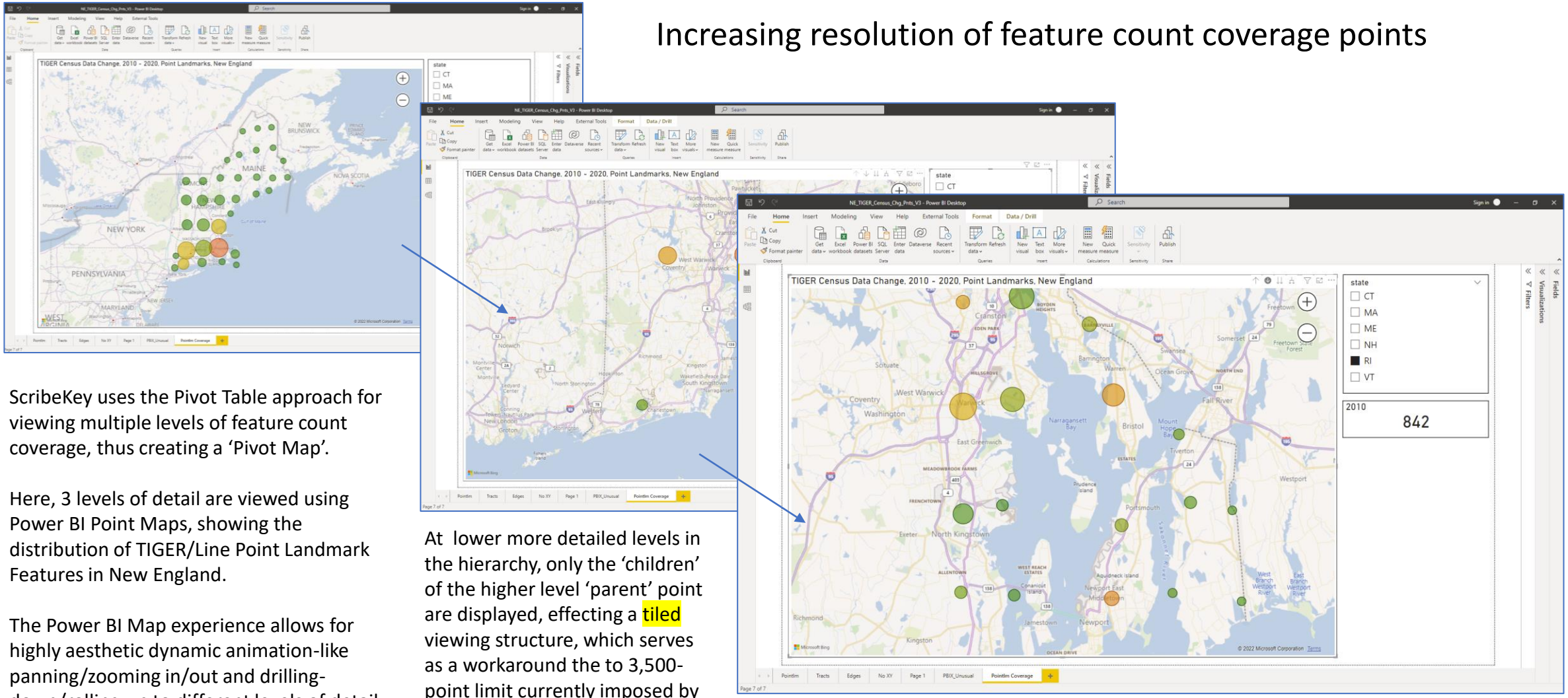
Here New England population is presented at 2 levels, state and county.



PivotTable: Rollup by State

Pivot Maps: 1 Degree, 1/10th Degree, 1/100th Degree

Increasing resolution of feature count coverage points



ScribeKey uses the Pivot Table approach for viewing multiple levels of feature count coverage, thus creating a 'Pivot Map'.

Here, 3 levels of detail are viewed using Power BI Point Maps, showing the distribution of TIGER/Line Point Landmark Features in New England.

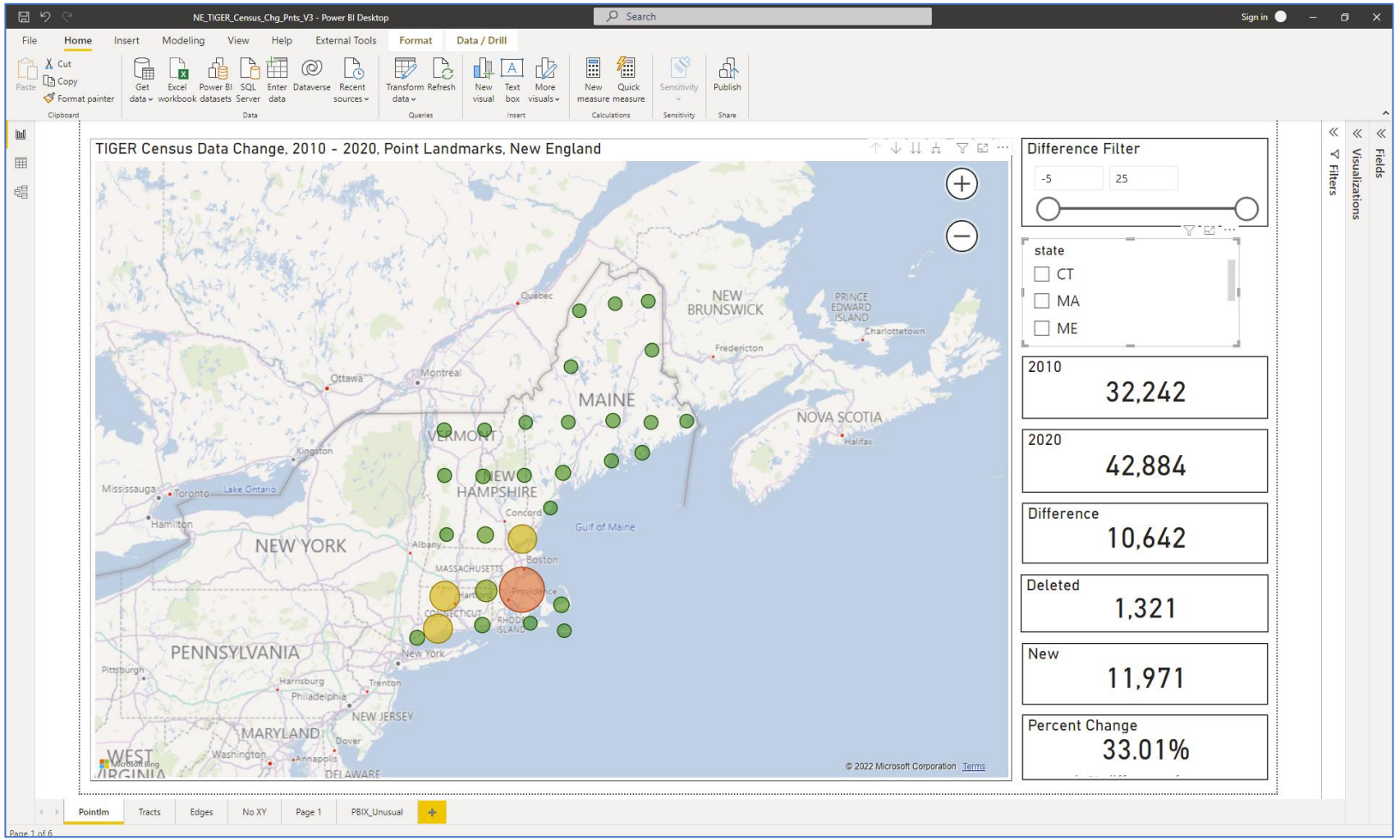
The Power BI Map experience allows for highly aesthetic dynamic animation-like panning/zooming in/out and drilling-down/rolling-up to different levels of detail.

At lower more detailed levels in the hierarchy, only the 'children' of the higher level 'parent' point are displayed, effecting a tiled viewing structure, which serves as a workaround the to 3,500-point limit currently imposed by the Power BI Map Point Visualization.

Change vs. Coverage

To complete the picture of ScribeKey's change metrics capture and visualization: Instead of using feature count coverage points, the amount of change in feature counts is attached to each Grid Map Point.

Additionally, metrics for the feature counts in each time period, the difference, the percentage change, and the number of new and deleted features, is available at each level in the Point Grid hierarchy.



TIGER Census Point Landmark Features, Change in Feature Count, 2010 - 2020

The Importance of Key Fields in the Data

TIGER Point Landmark features have a permanent unique primary key field called POINTID. Edges and Tracts also have unique identifiers.

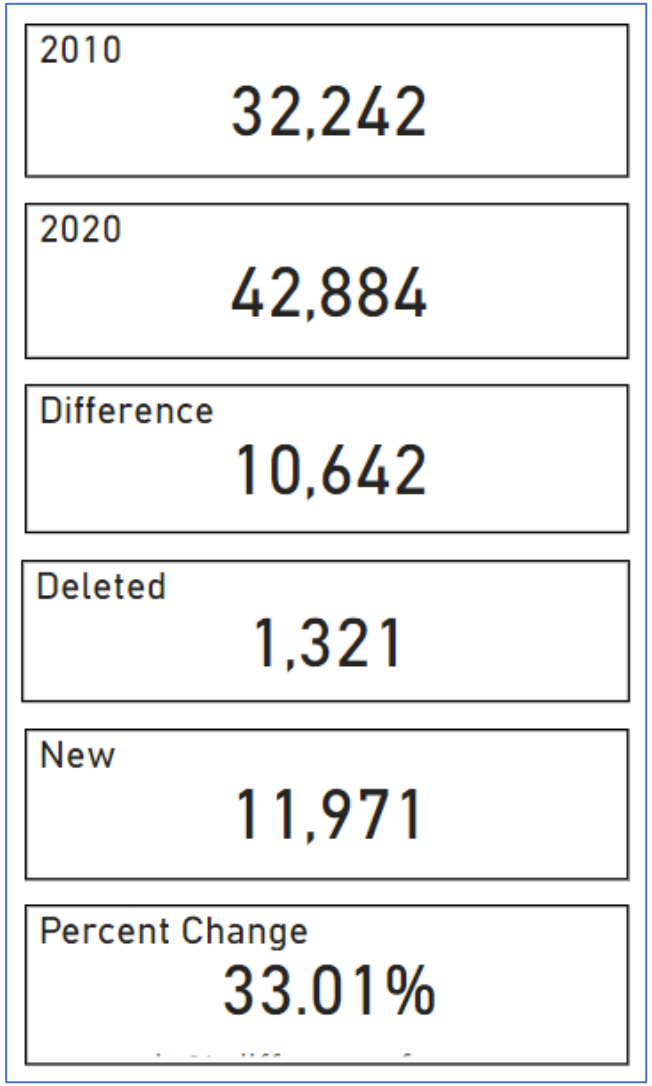
Features are not duplicated in this data.

It is more work for the Census to maintain this as permanent, but it allows for exact comparisons between datasets of different time periods.

It allows for the determination of which features are the same in both time periods, which features have been deleted, and which features are new.

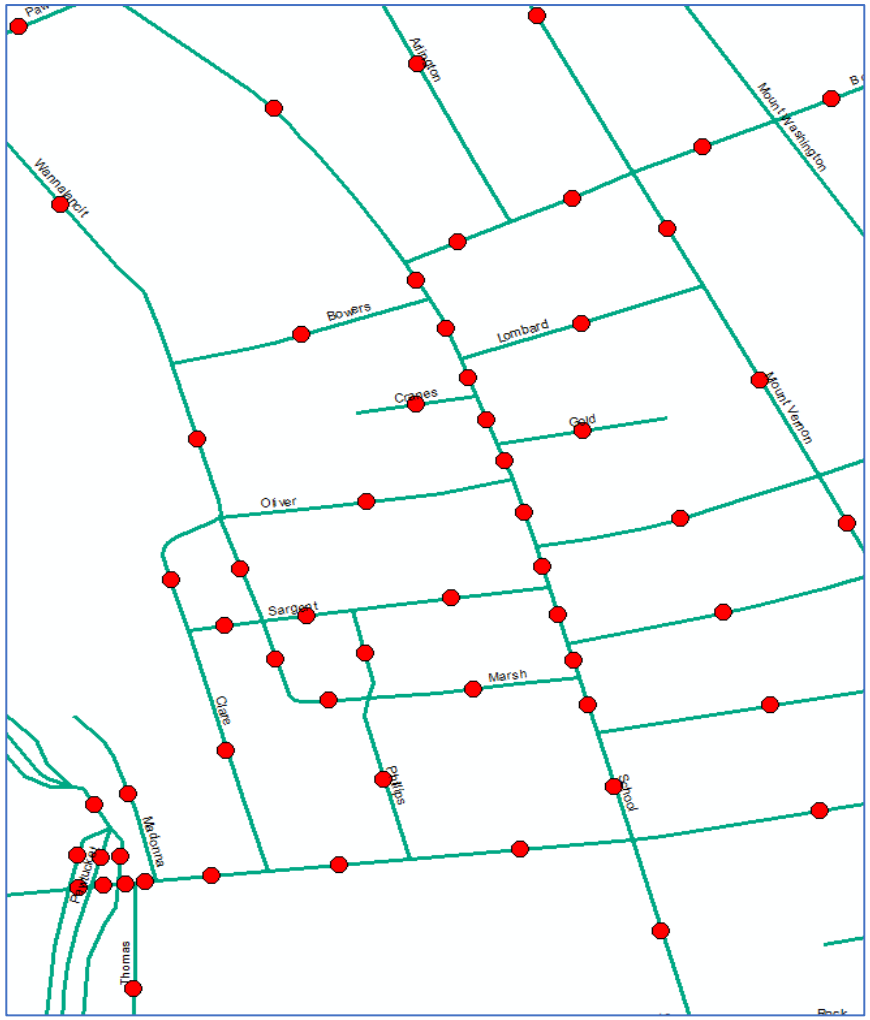
Not all data preserves unique features by key. For example, HERE Technologies (formerly Navteq), sometimes stacks multiple features with the same key in the same location because of language differences and name variants.

When this occurs, same, deleted, and new features need to be counted differently.



Geometry Details: Adding XY

- At the onset, in order to generate the aggregated Grid Map Points, XY Centroid values are added to all features in the layers being analyzed.
- XY values are also added to non-geometric tables, e.g., ADDR when a common key, e.g., TFIDL is available. This supports seeing ‘where’ table rows have changed.
- Faster alternate to actual centroid is Center of range box – incurs some aggregation generalization but is much faster.
- This allows for a generalized spatial comparison between all features and table rows common to both time periods.
- While generalized, centroid location is highly sensitive to any geometry changes.
- Adding XY to NE Edges with ArcGIS Pro 64-bit python: 130 seconds.
- Adding XY to NE Edges with Amazon’s Redshift: 10 seconds.
- ESRI Arcpy NEAR () function can also produce results very quickly.
- This XY work is fundamental to the creation of hierarchical change points as well as being used for Geometry Updates analysis.



XY Centroids added to Street Edges

Create the Base Change Points Data Table

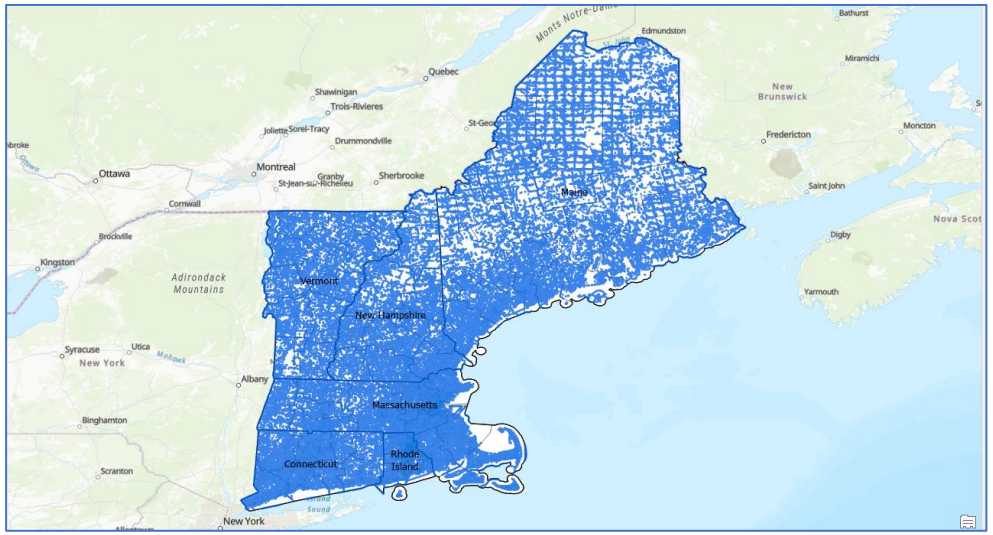
- XY points are generalized, ‘snapped to’ the closest point grid features.
- 1 degree, 1/10th degree, 1/100th degree.
- Identifier data is also included, layer name, state, county, etc.
- A single 100th degree XY table is created for Power BI, multiple hierarchical layers need to be created for ArcGIS or QGIS.
- **Fundamental: After starting with high level feature count differences from data profiling, these change points allow us to see exactly WHERE change has occurred at multiple levels.**

lyr	a	b	same	del	new	diff	x100	y100	x10	y10	x1	y1
edges	1	1	1	0	0	0	-68.47	45.41	-68.5	45.4	-68	45
edges	1	1	1	0	0	0	-68.47	45.44	-68.5	45.4	-68	45
edges	1	1	1	0	0	0	-68.45	45.39	-68.5	45.4	-68	45
edges	1	1	1	0	0	0	-68.44	45.35	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.44	45.4	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.43	45.37	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.43	45.42	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.42	45.37	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.4	45.36	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.4	45.43	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.4	45.44	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.38	45.37	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.38	45.38	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.38	45.43	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.37	45.35	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.36	45.39	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.36	45.42	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.36	45.43	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.35	45.37	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.35	45.38	-68.4	45.4	-68	45
edges	1	1	1	0	0	0	-68.34	45.43	-68.3	45.4	-68	45
edges	1	1	1	0	0	0	-68.33	45.39	-68.3	45.4	-68	45

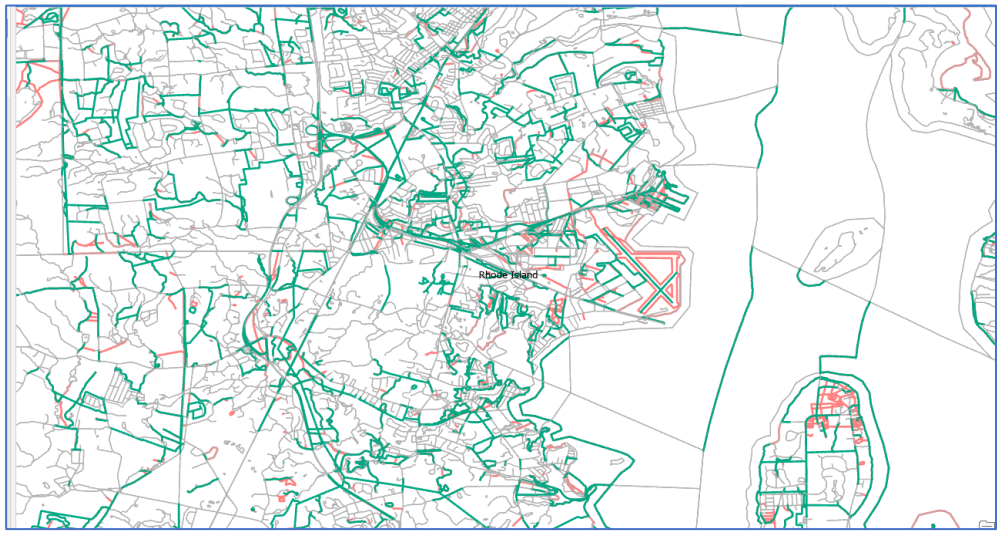
Single XY 100th Degree Data Change Table
TIGER Edges Feature Points - Power BI View

ArcGIS: Feature Detail 1/3

- Power BI is a good tool for smaller sized point feature grid datasets and hierarchical viewing.
- However, Power BI is not suited to display millions of geometric features.
- For this kind of detail, a GIS platform like ArcGIS or QGIS is required.
- ‘Same’ features are those common to both time periods, Deleted have been removed from the older dataset, New have been added in the most recent dataset.

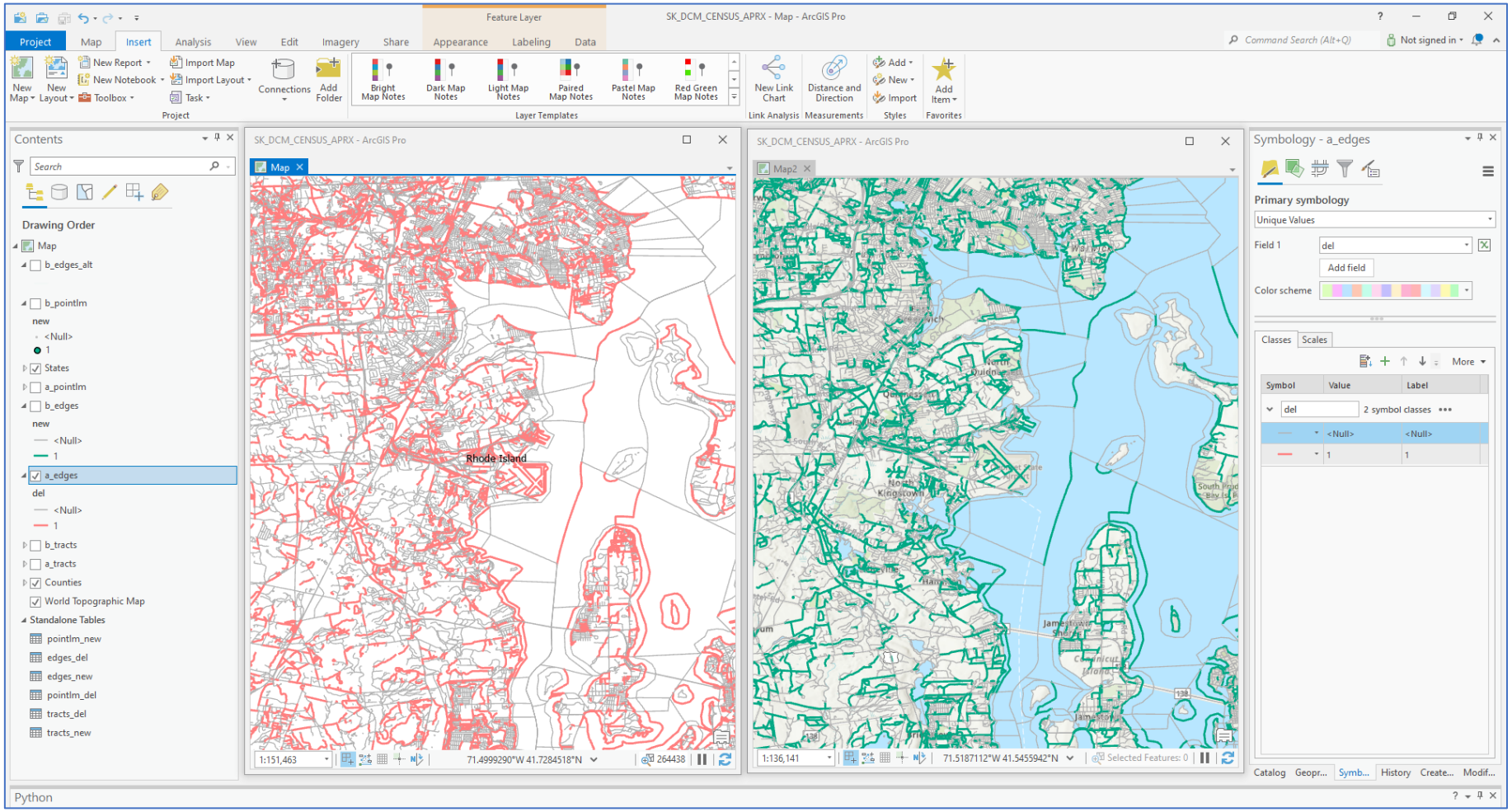


New Edges Features Selected in ArcGIS Pro



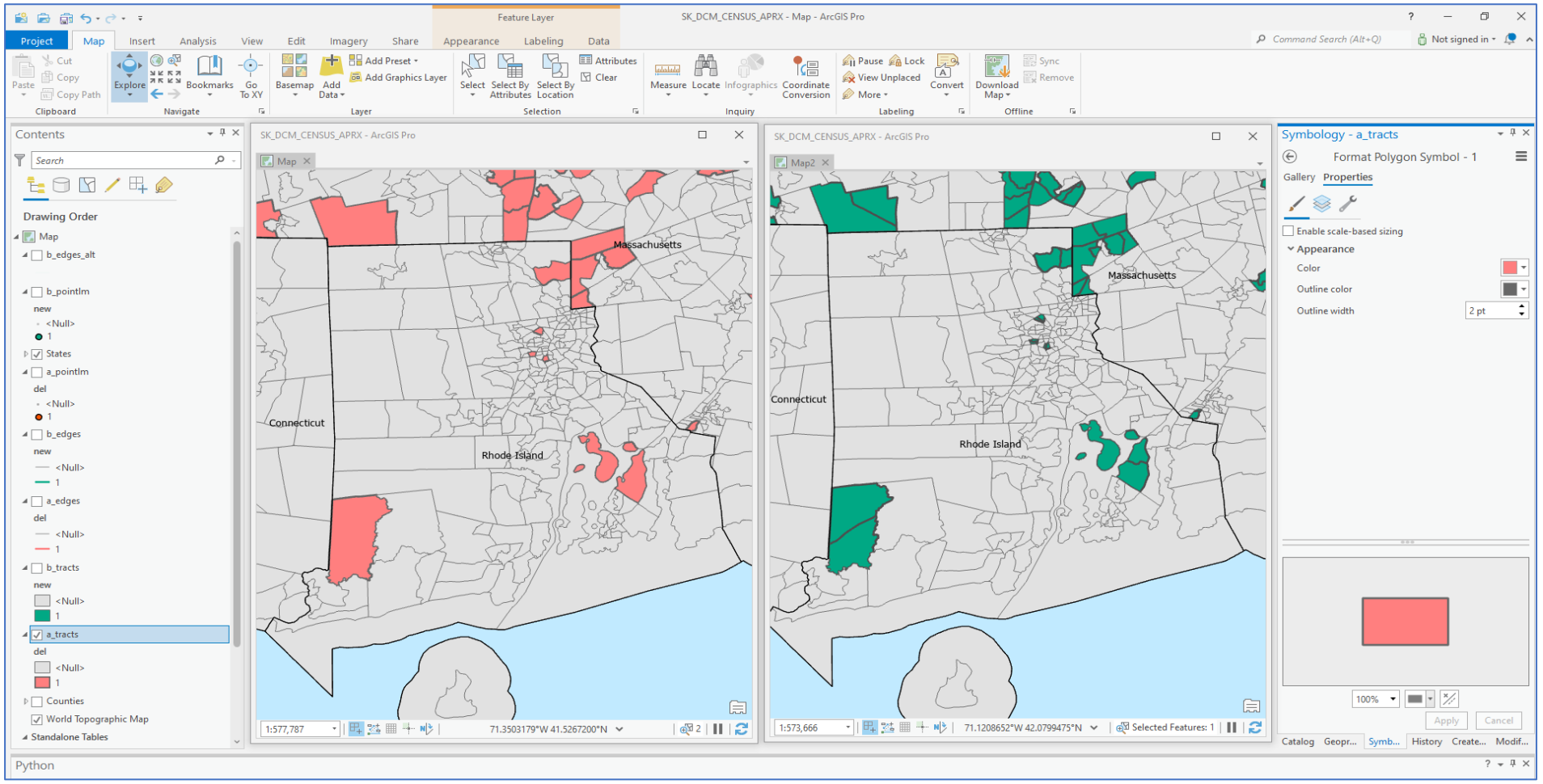
Same Grey, Deleted Red, New Green Feature Detail

ArcGIS: Feature Detail 2/3



Data Change Cartography: Side by Side vs Overlay
TIGER/Line Edges, Deleted and New, in Rhode Island

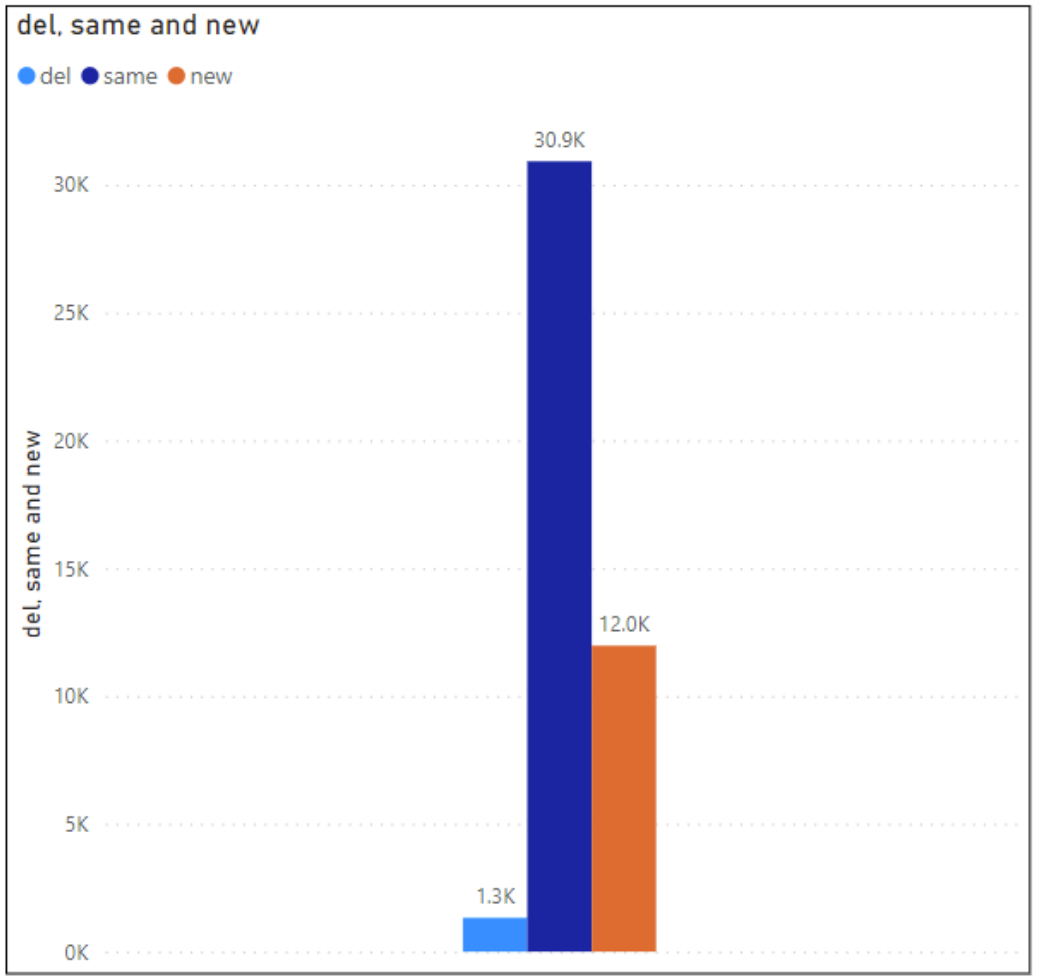
ArcGIS: Feature Detail 3/3



Changes in Census Tracts, because of smaller data volume, are much easier to see.

Churn: Attribute and Geometry Updates

- Up to this point, we have reviewed the delta in feature counts along with counts of deleted and new features.
- Additionally, for the features common to both time periods (‘same’) we want to capture and present attribute value and geometry changes.
- We also want to apply the Point Grid Map change points technology to these ‘churn’ metrics to see exactly WHERE the changes have occurred.
- The bar chart on the right is typical for showing relative quantities of Deleted, Same, and New Features.

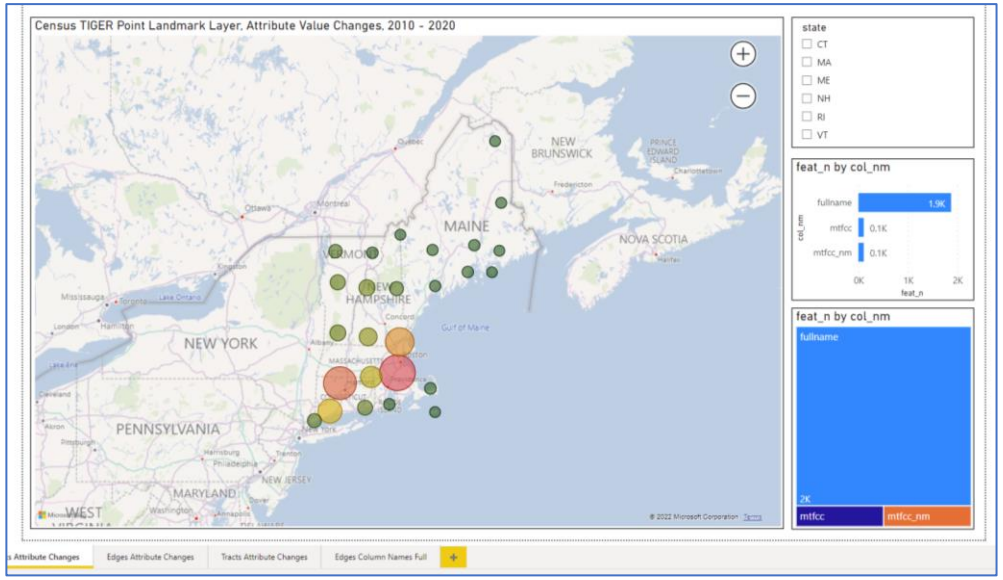


Deleted, Same, and New POINTLM Features in Power BI

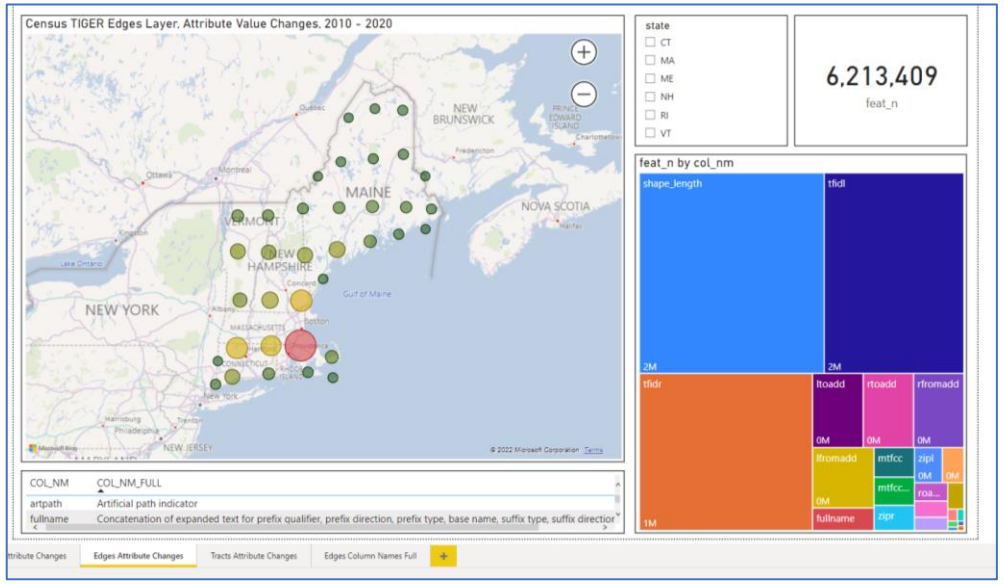
Attribute Value Updates

- The distribution and clustering of attribute value changes across the 3 layers is similar, but still shows some differences.

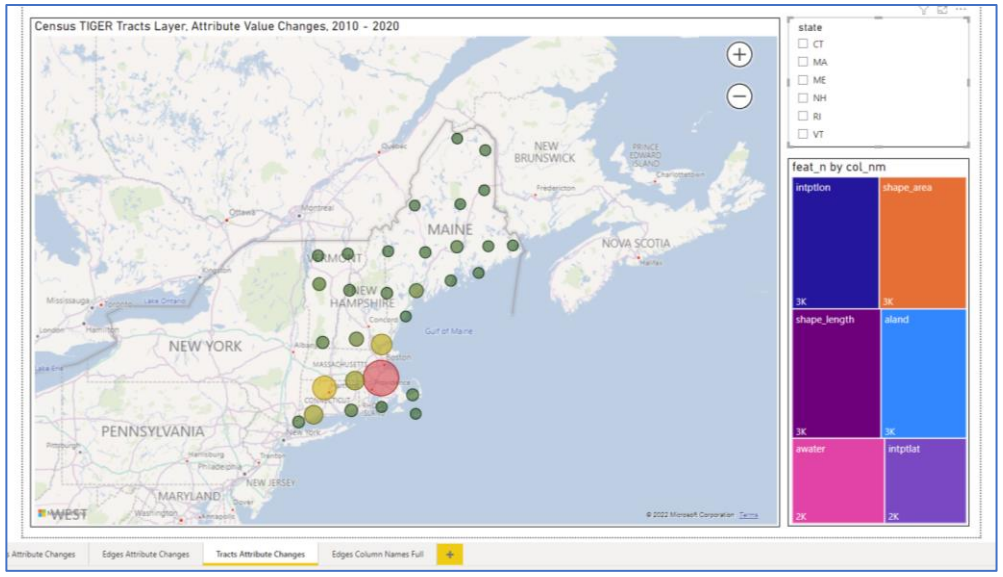
- We can again, drill down to areas of concentrated change, using hierarchical change point maps in MS Power BI.



Point Landmarks



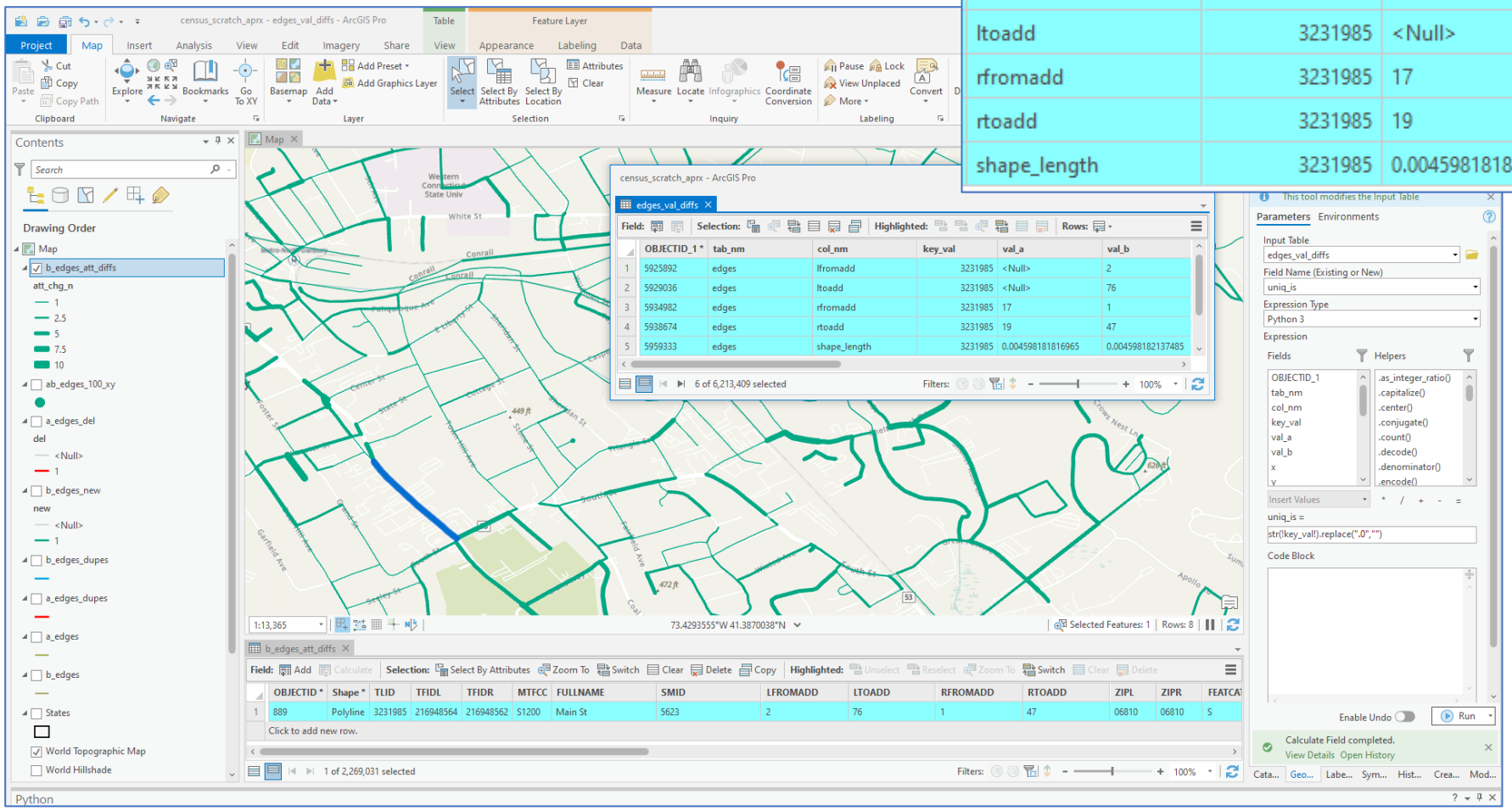
Edges



Census Tracts

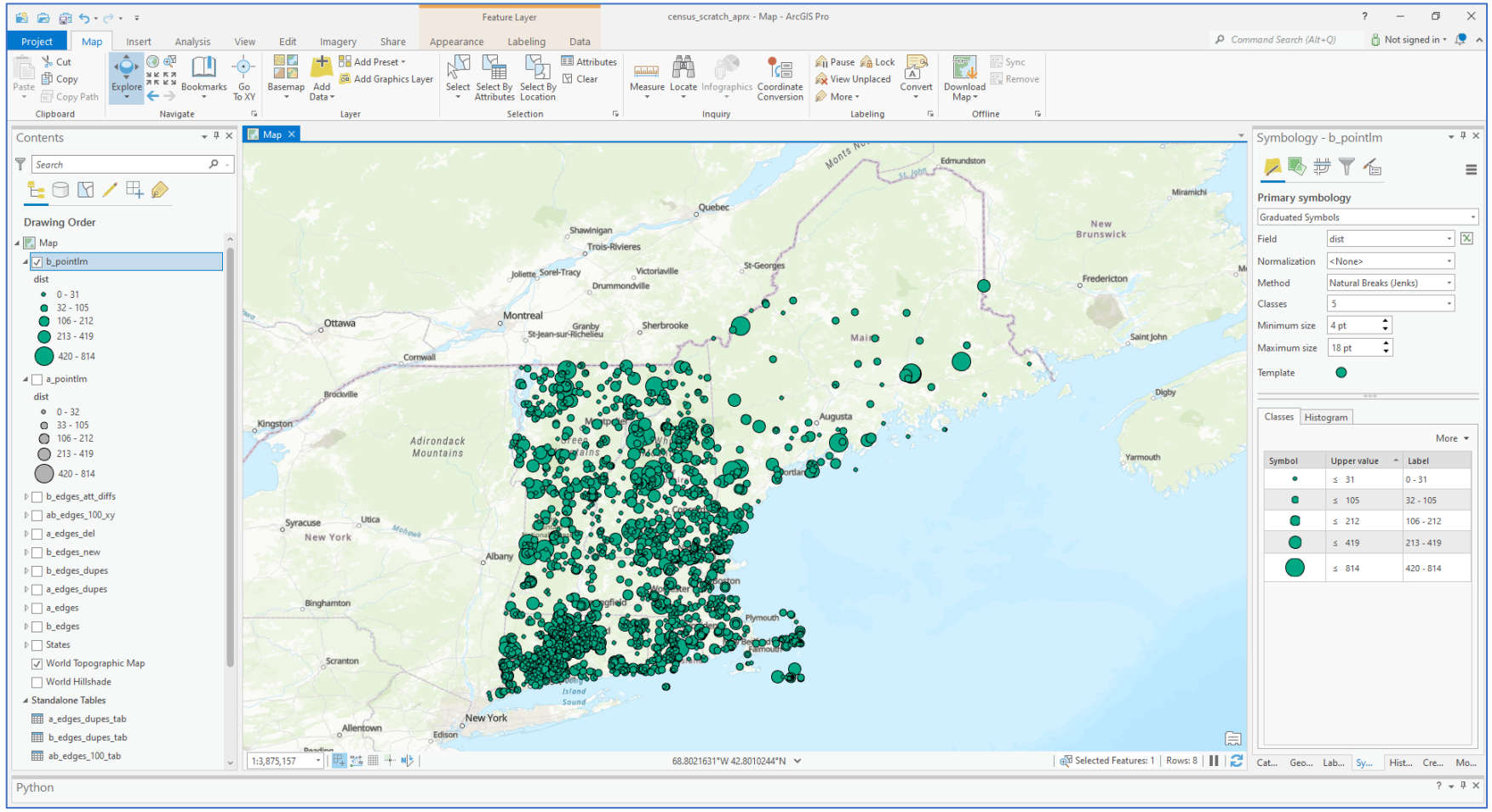
Attribute Value Updates in ArcGIS

col_nm	key_val	val_a	val_b
lfromadd	3231985	<Null>	2
ltoadd	3231985	<Null>	76
rfromadd	3231985	17	1
rtoadd	3231985	19	47
shape_length	3231985	0.004598181816965	0.004598182137485



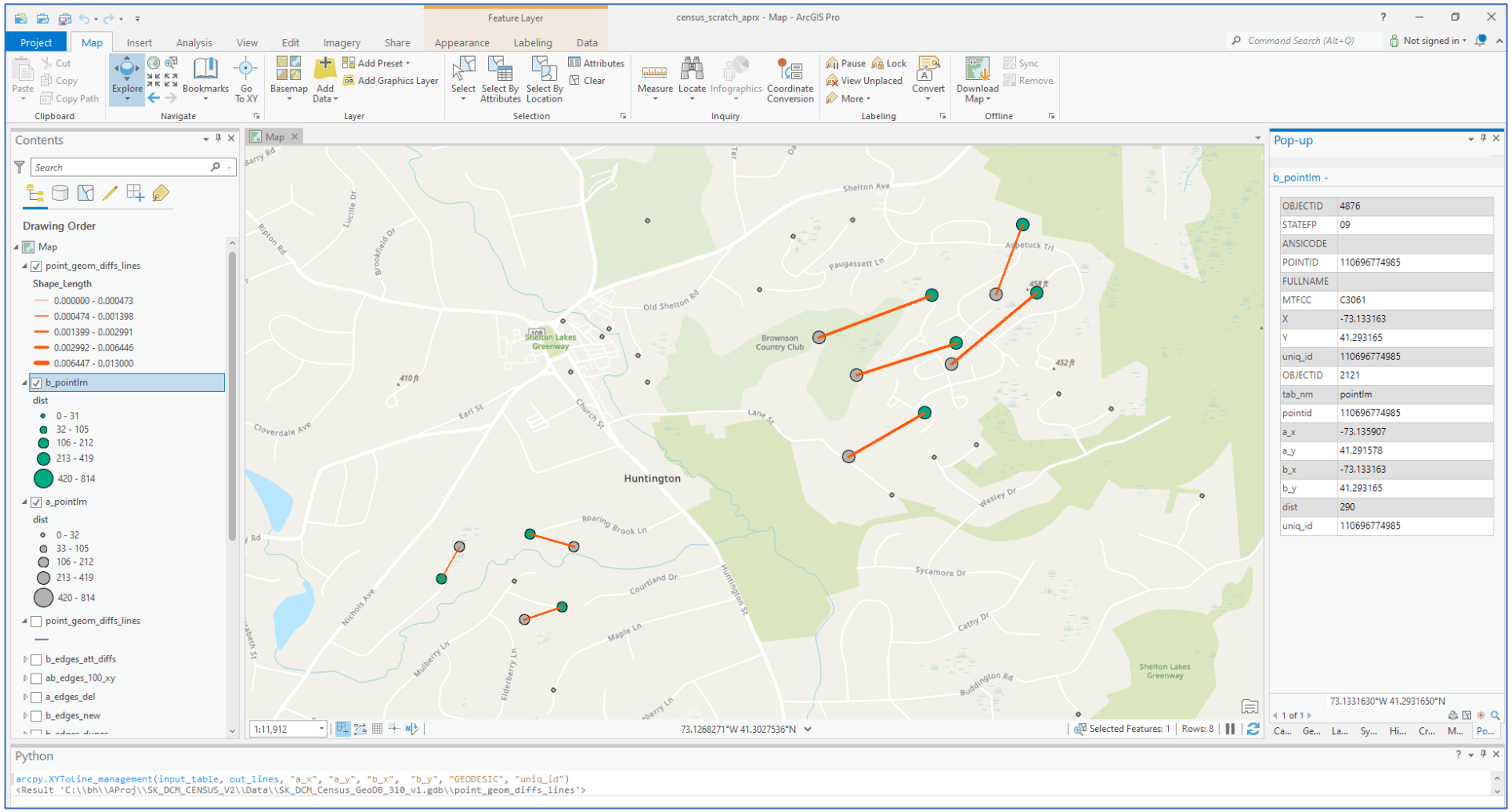
Using ArcGIS RELATE functionality, the list of all attribute edits can be reviewed on an individual feature basis. Attribute value changes comprise the highest volume change metrics data, with millions of rows.

Geometry Updates



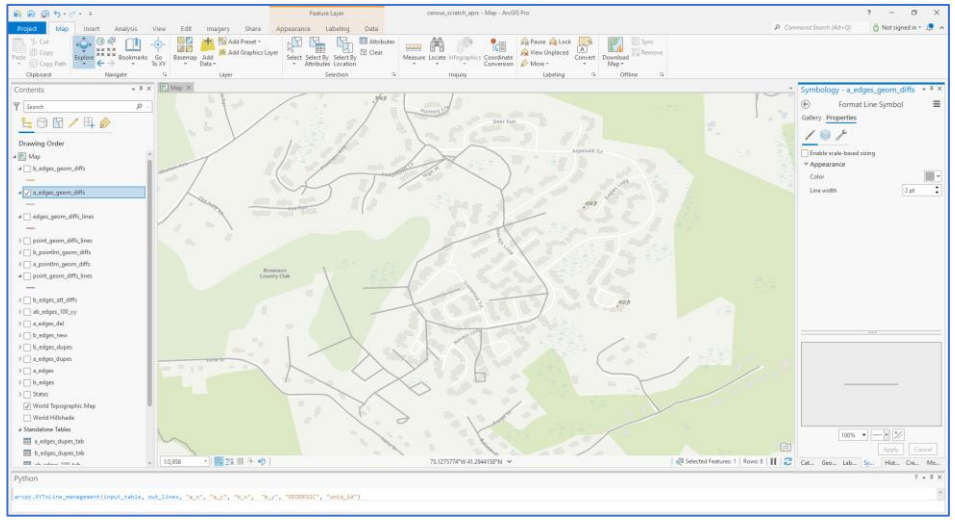
The distance a feature has moved from the older dataset to the newer dataset can also be captured. Here, the size of a Point Landmark feature is scaled by the distance it has been moved in the time period between the 2 datasets, 2010 and 2020.

Geometry Updates: Distance Moved Lines Added

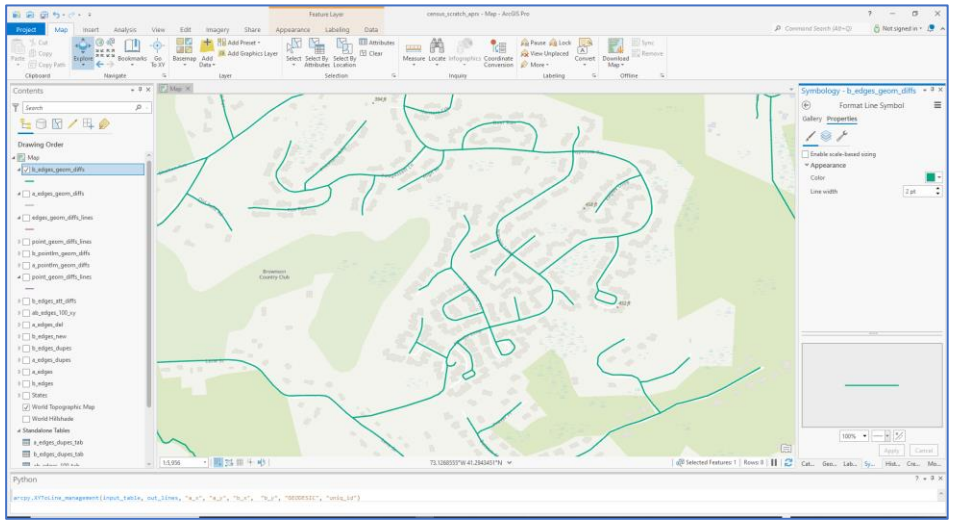


Line features from the older location to the newer location are generated alongside Point Landmark Features. Older features shown in grey, newer in green.

TIGER Edges Layer Polyline Geometry Updates

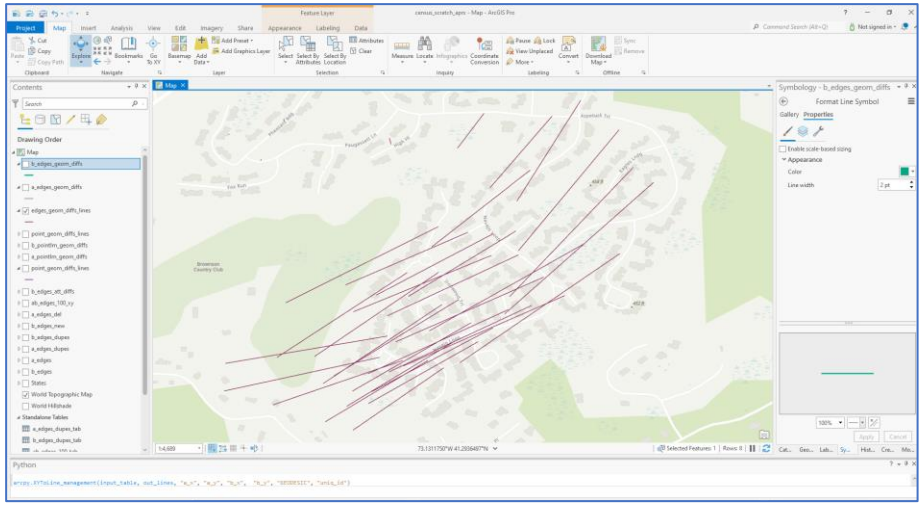


2010



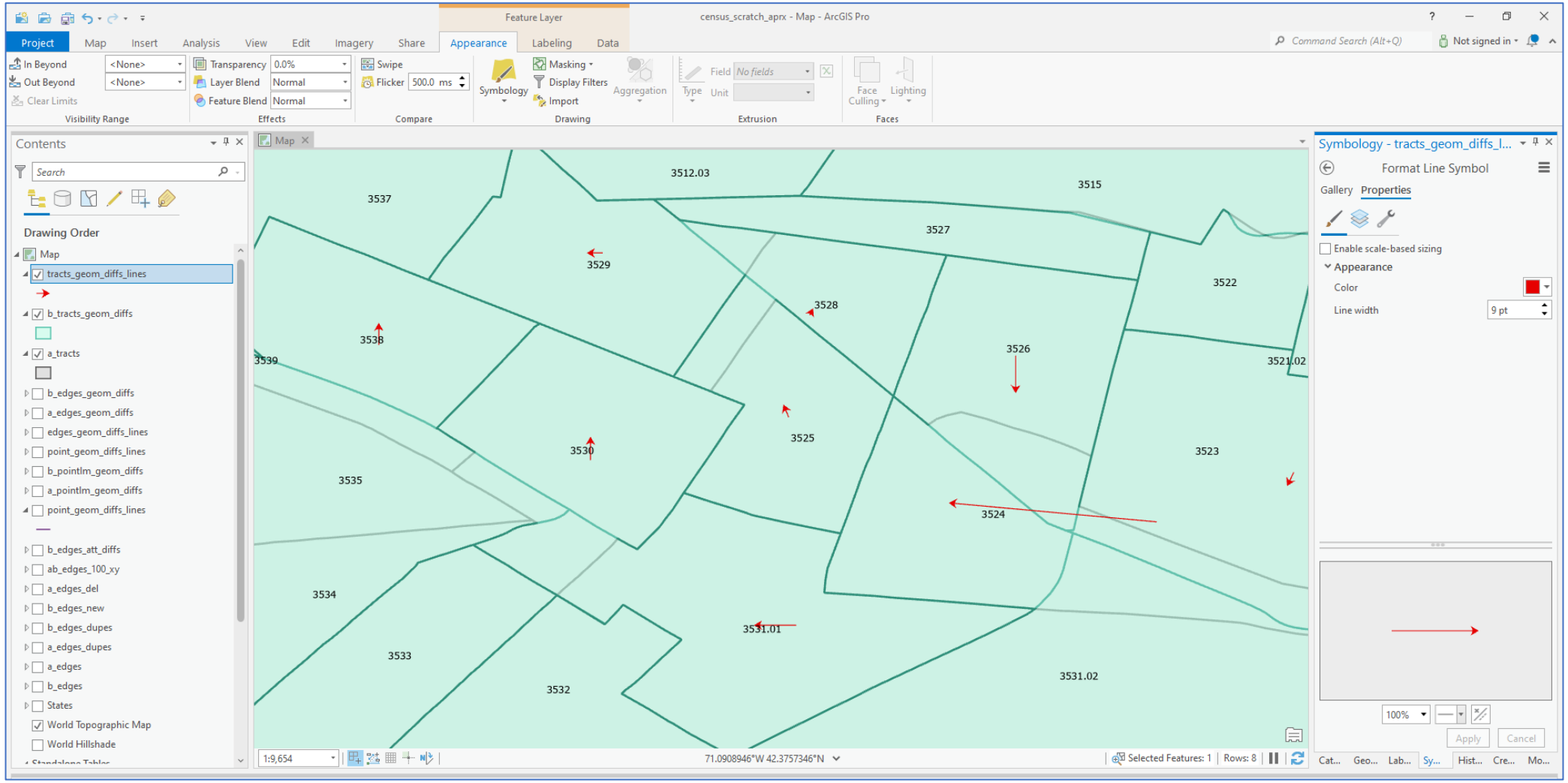
2020

Polyline and Polygon Geometry Updates are Captured as Changes in the Location of a Feature's Centroid. While not exact, it is highly sensitive to change.



Edges Centroid Movements Over the Time Period

Updates to Census Tract Polygons



Older and Newer Census Tract Polygons, Superimposed, Different Colors, Distance Movement Arrows

Including Metadata in Attribute Change Metrics

It is particularly crucial to understand what the data fields mean when studying attribute value updates.

Here, data dictionary elements from TIGER/Line data documentation are added to the Power BI Report file.

COL_NM	COL_NM_FULL
artpath	Artificial path indicator
fullname	Concatenation of expanded text for prefix qualifier, prefix direction, prefix type, base name, suffix type, suffix direction, and suffix qualifier with a space between each expanc (as available)
deckedroad	Decked road indicator
divroad	Divided road
shape_length	ESRI geometry length
exttyp	Extension type
lfromadd	From house number associated with the most inclusive address range on the left side of the edge
rfromadd	From house number associated with the most inclusive address range on the right side of the edge
tnidf	From TIGER node identifier
featcat	General feature classification category
persist	Hydrographic persistence flag
hydroflg	Hydrography feature indicator
offsetl	Left offset flag
mtfcc	MAF/TIGER Feature Class Code of the primary feature for the edge
mtfcc_nm	MTFCC Full name
olfflg	Other linear feature indicator
railflg	Rail feature indicator
offsetr	Right offset flag
roadflg	Road feature indicator
gcseflg	Short lines flag for geographic corridors
mtfcc_cat	Single char MTFCC category
smid	Spatial metadata identifier
passflg	Special passage flag
ltoadd	To house number associated with the most inclusive address range on the left side of the edge
rtoadd	To house number associated with the most inclusive address range on the right side of the edge
tnidt	To TIGER node identifier
tfidl	Topological Faces Identifier on the left side of the edge
tfidr	Topological Faces Identifier on the right side of the edge
ttyp	Track type
zipl	ZIP Code associated with the most inclusive address range on the left side
zipr	ZIP Code associated with the most inclusive address range on the right side

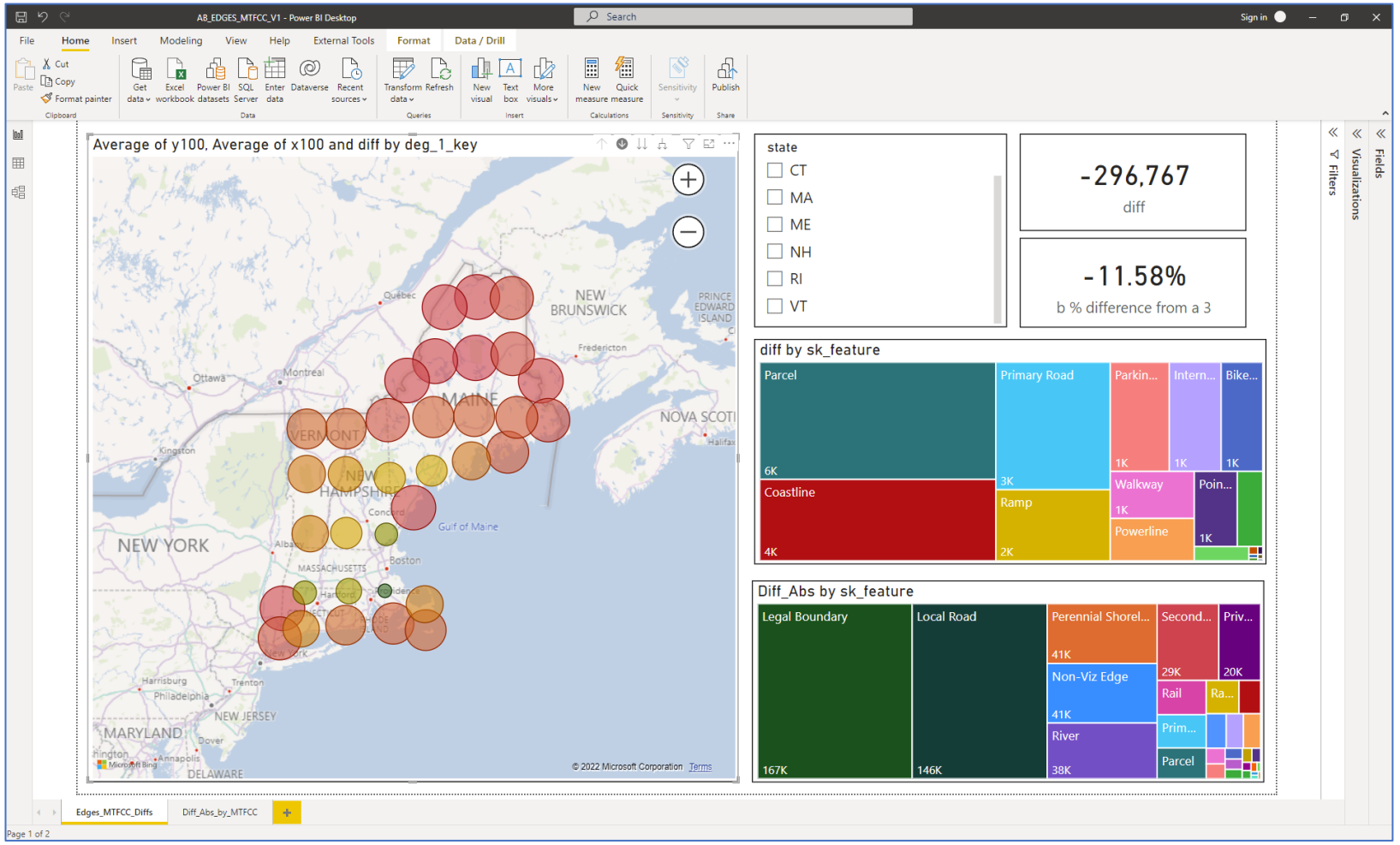
TIGER Census Edges Column Abbreviations and Full Names
Captured from Census Metadata Documents

More Granularity With Feature Subtypes: MTFCC

We can increase the granularity of our comparisons by adding a feature classification identifier, MTFCC, into the aggregated change points tables.

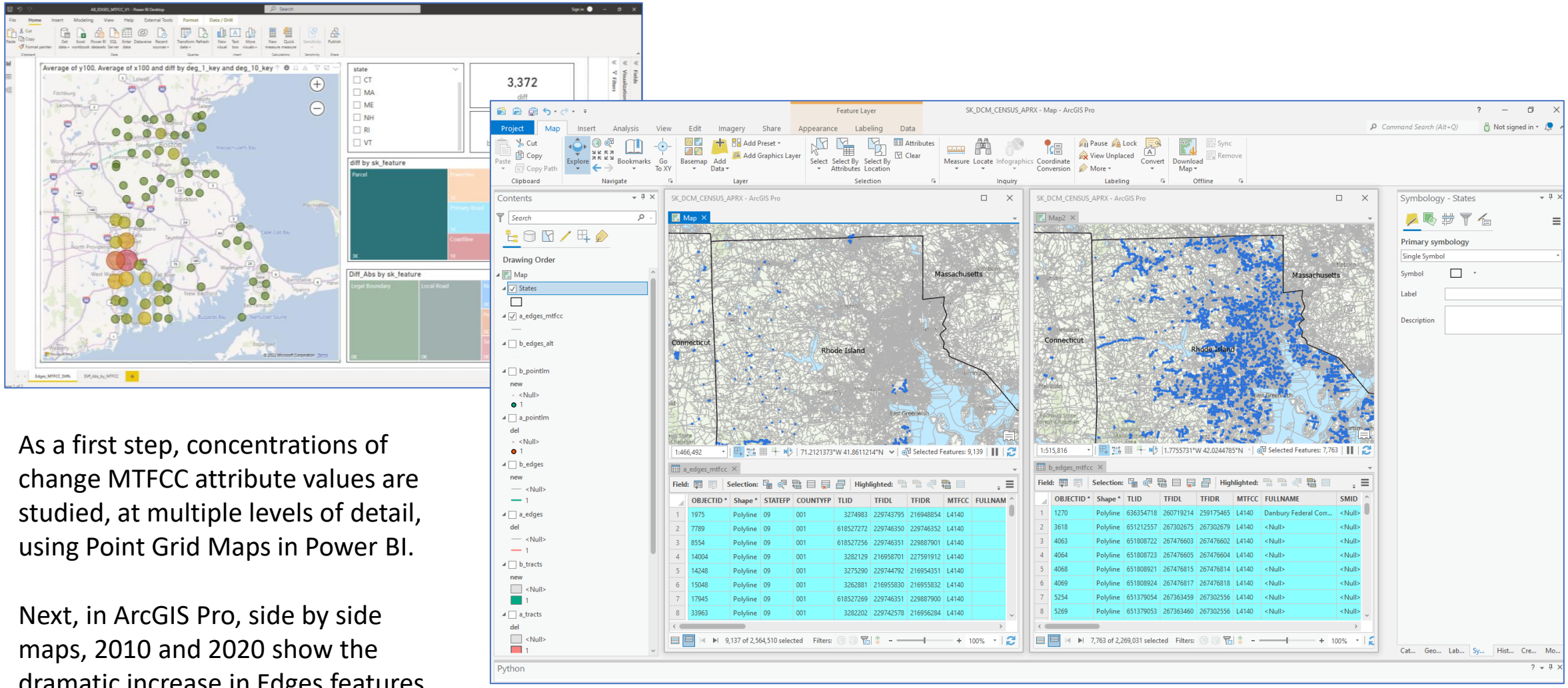
We can see significant increases in Parcel and Coastline features, and a decrease in Legal Boundary and Local Road features.

Next, we use Power BI to see WHERE these differences occur.



Changes in Feature Count by MTFCC with Tree Diagrams, including ABS()

More Parcels in Rhode Island

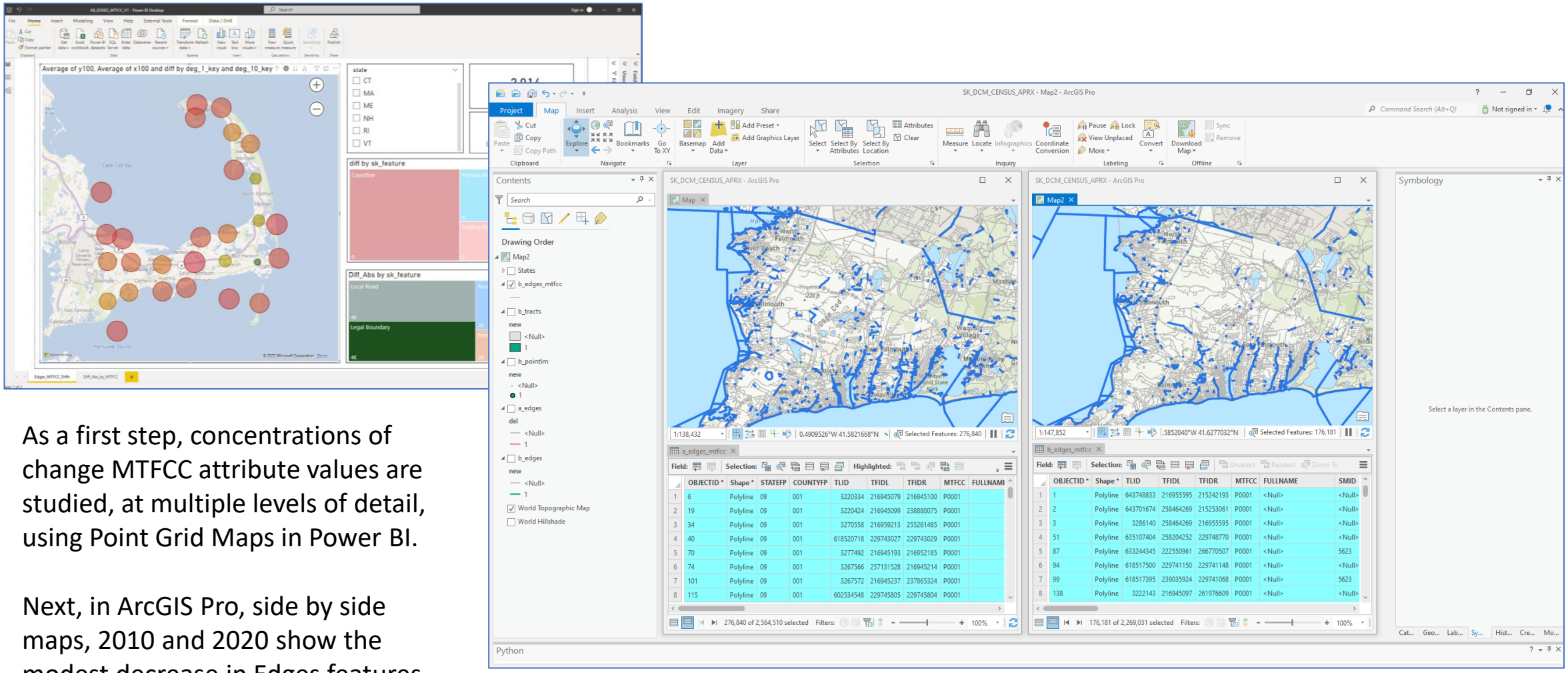


As a first step, concentrations of change MTFCC attribute values are studied, at multiple levels of detail, using Point Grid Maps in Power BI.

Next, in ArcGIS Pro, side by side maps, 2010 and 2020 show the dramatic increase in Edges features corresponding to Parcel Lines.

Increase in Parcel Lines MTFCC Feature Types in Rhode Island

Fewer Legal Boundaries, Somewhat Widespread
Here Example is on Cape Cod



As a first step, concentrations of change MTFCC attribute values are studied, at multiple levels of detail, using Point Grid Maps in Power BI.

Next, in ArcGIS Pro, side by side maps, 2010 and 2020 show the modest decrease in Edges features corresponding to Legal Boundaries.

Decrease in Legal Boundary MTFCC Feature Types on Cape Cod

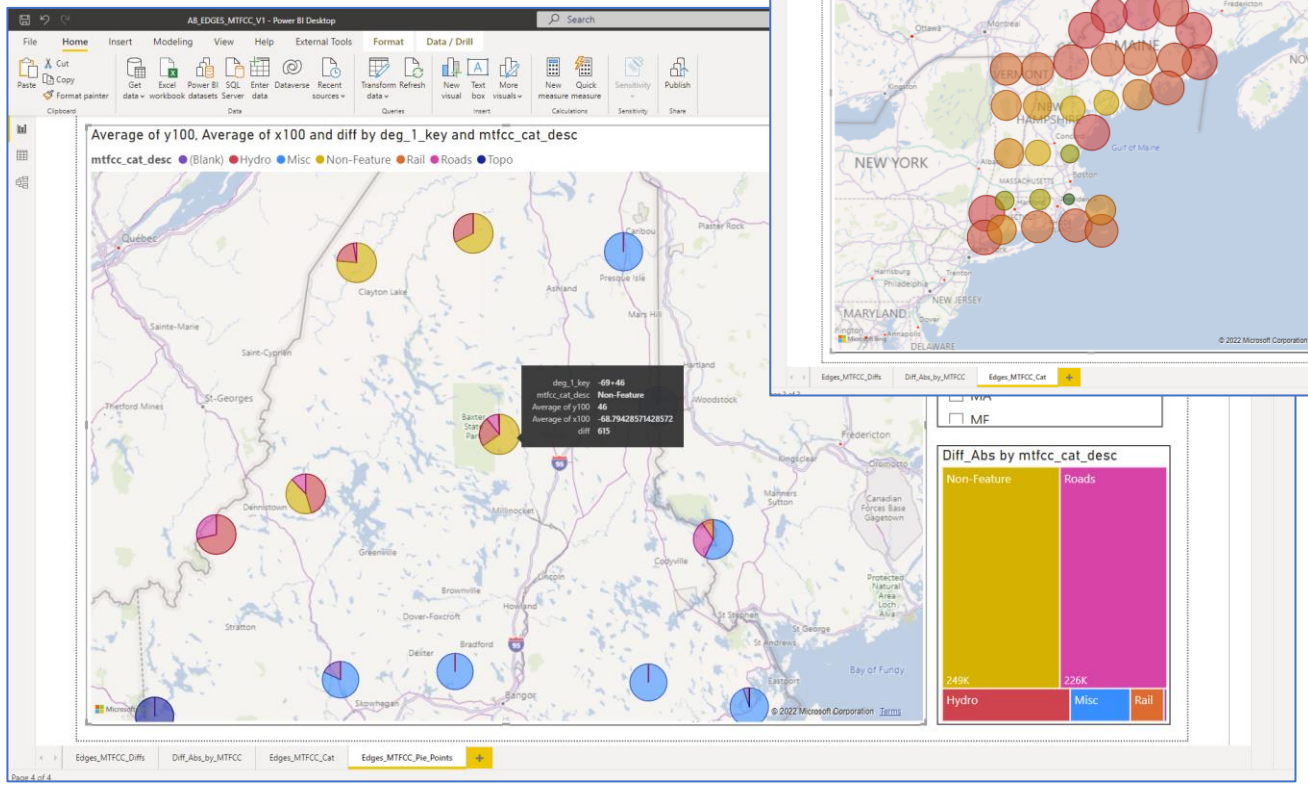
More General MTFCC Aggregation: 6 Categories

MTFCC_CAT	DESC
H	Hydro
L	Misc
S	Roads
C	Topo
R	Rail
P	Non-Feature

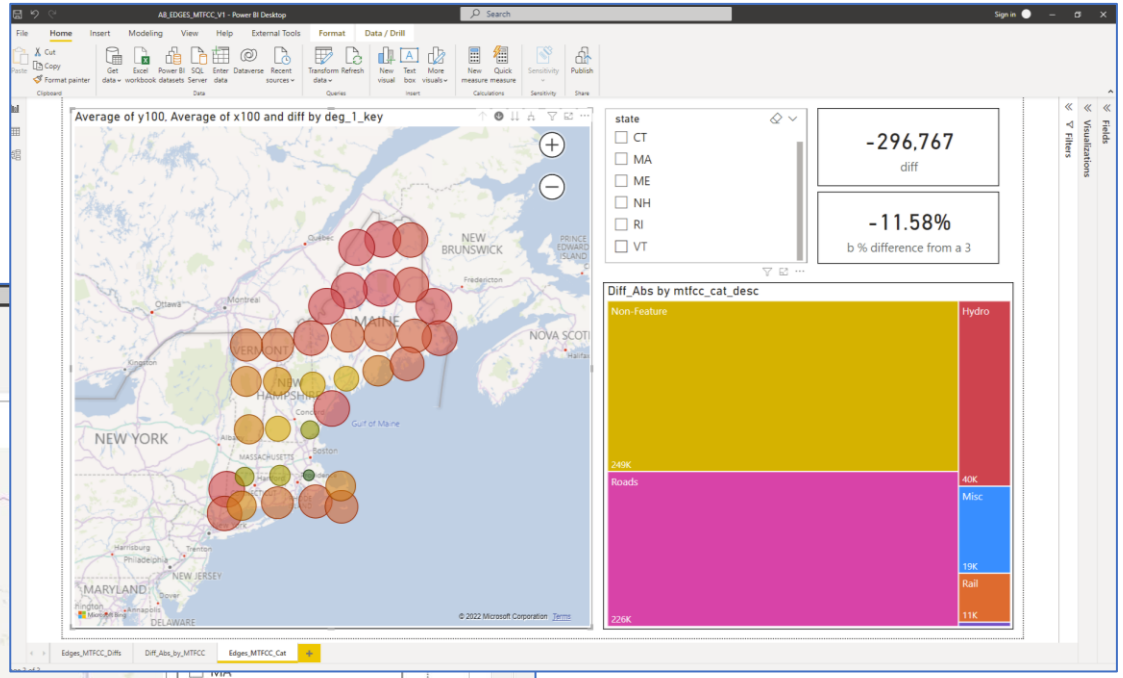
MTFCC Categories, as the first letter in the MTFCC code string.

Yet another level of aggregation using more generalized MTFCC Categories.

6 vs. 39 Feature Type Values.



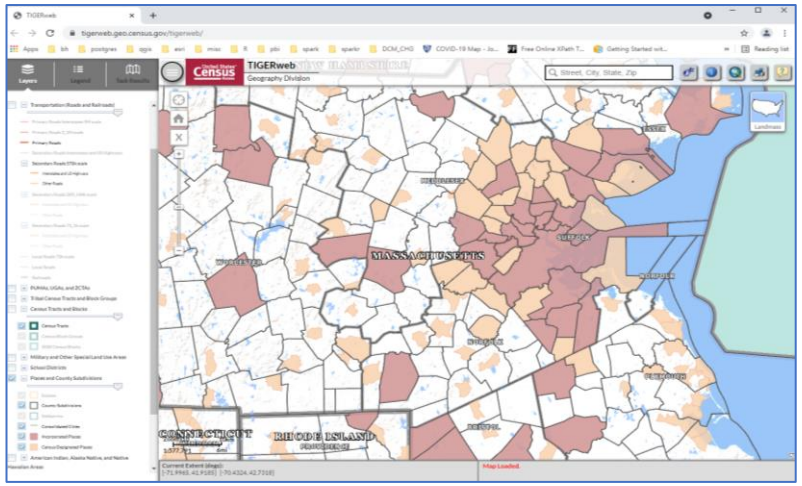
Power BI Legend = MTFCC_Cat Pie Chart Points



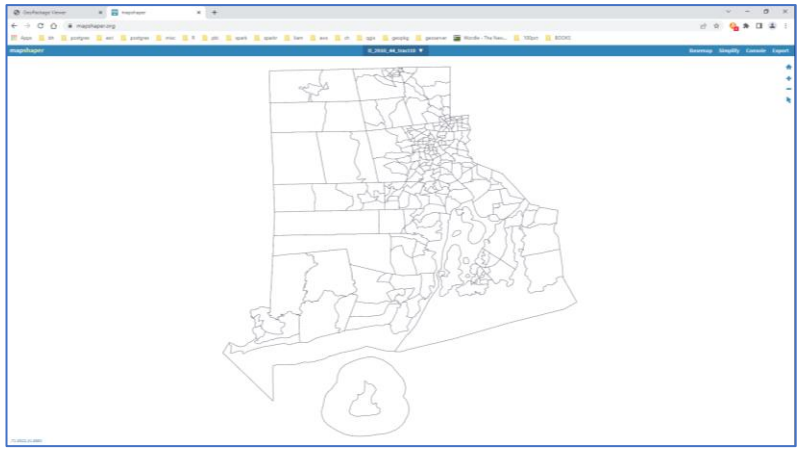
Power BI - No Legend

User Friendly Geography, Polygons, Spatial Join

- When creating hierarchical geographic datasets, the obvious first inclination is to use administrative polygons, e.g., State, County, Town.
- These are not at all abstract, familiar to everyone, and easily lend themselves to BI pivot table style analysis.
- But for large volume vector datasets, not already tagged with the polygons enclosing them, expensive spatial joins are required.
- The hierarchical grid point technique described in this presentation is much faster to create and use in a tool like Power BI.
- The hierarchical grid points can also be joined to administrative polygons much faster than the original underlying features.
- Mapshaper.org is a great site for converting highly detailed administrative polygons to simpler more generalized features, which display much faster.



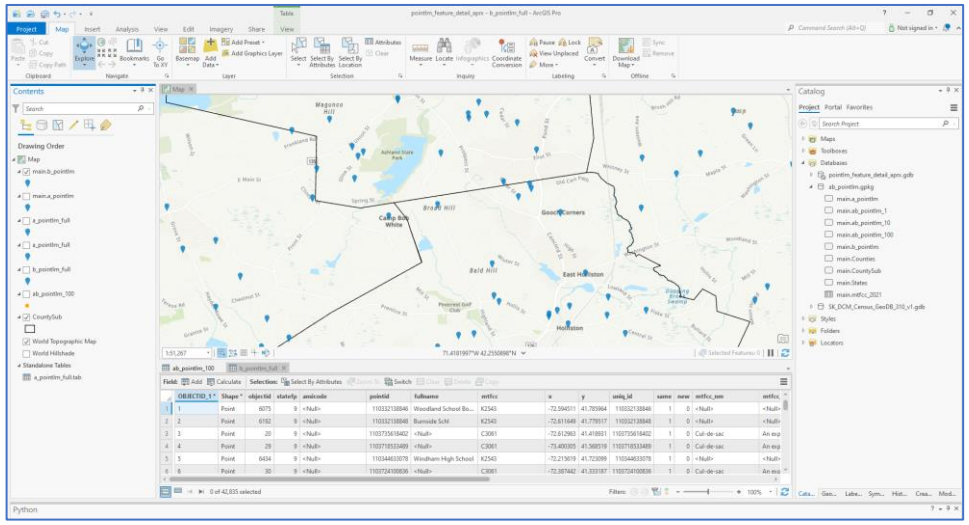
TIGERweb



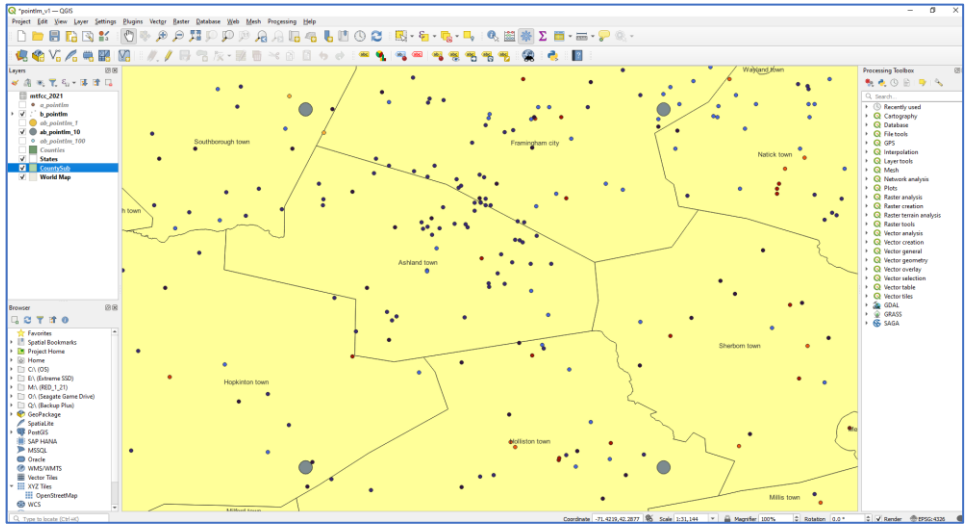
https://mapshaper.org/

The Geopackage

- In forthcoming efforts, ScribeKey will be capturing data coverage and data change using the Geopackage.
- Geopackages: Open source, based on Spatialite (Sqlite - a misnomer, can handle very large datasets)
- Vector, raster, tables, metadata.
- Full SQL with spatial functions.
- Single file but can query/join across multiple files
- VERY fast and flexible
- Use/edit with QGIS, ESRI, GDAL, Python, R, SQLite and Postgres
- QGIS Python (Plugins), Command Line Interface (CLI), Other Python libraries
- Enjoying increasing use, related to QGIS, which also runs on Mac



TIGER 2020 Point Landmarks, Geopackage, in ArcGIS Pro



TIGER 2020 Point Landmarks, Geopackage, in QGIS

Requirements and Environments

- Goal: Capture and present comprehensive data change information in high volume multi-layer/table vector datasets.
- At the row level show features common between 2 snapshots in time, as well as deleted and new features, relying on permanent primary keys in the data.
- Show fully detailed data describing all attribute value updates.
- Show fully detailed data describing all geometry updates.
- Link non-geometry tables to feature tables through unique or stacks of primary keys and show these changes geospatially.
- Provide an environment for ad-hoc exploration of geospatial data change in both a BI environment (Power BI, Tableau) and a GIS environment (ArcGIS, QGIS) in both a local on-premise desktop setting and on the web.
- Generate systematic geospatial data change atlas-like bookmarks in both the BI and GIS environments (map books).
- In addition to desktop and web environments Big Data Engines like Amazon Web Service (AWS) Redshift and Athena, support processing of very high-volume datasets, while also supporting geometry data types.

- ArcGIS, ArcGIS Pro Desktop
- ArcGIS Online
- QGIS
- QGIS Cloud
- Power BI Desktop
- Power BI Service (web)
- Tableau Desktop
- Tableau Online
- Amazon Redshift
- Amazon Athena

There is a fundamental paradigm/technology divide between GIS and BI platforms. GIS doesn't provide pivot tables; BI doesn't provide support for high volume vector features.

Key Take Away

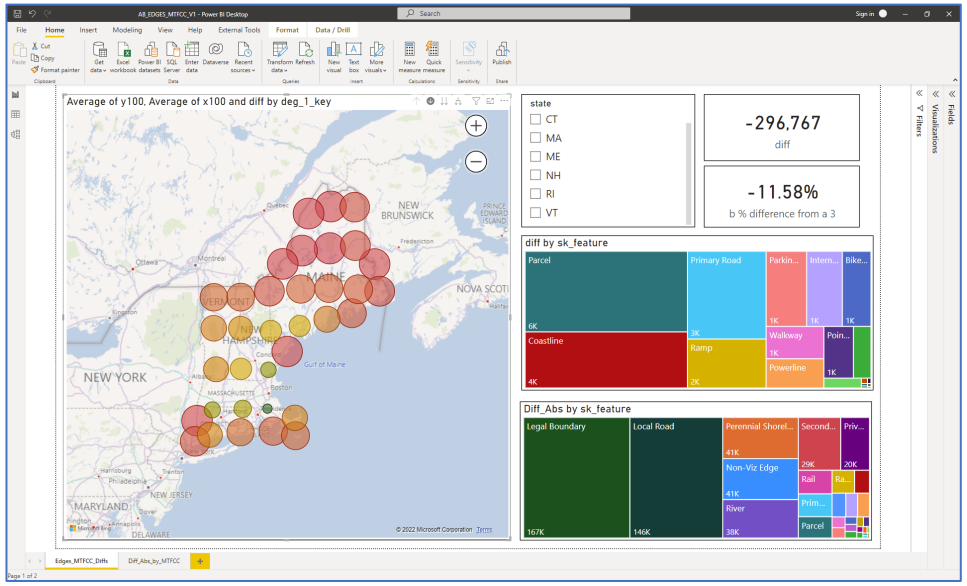
The essential nature of BI involves measures (numbers) and dimensions (categories). For example, show me the relative quantities of different Census TIGER MTFCC codes in a dataset.

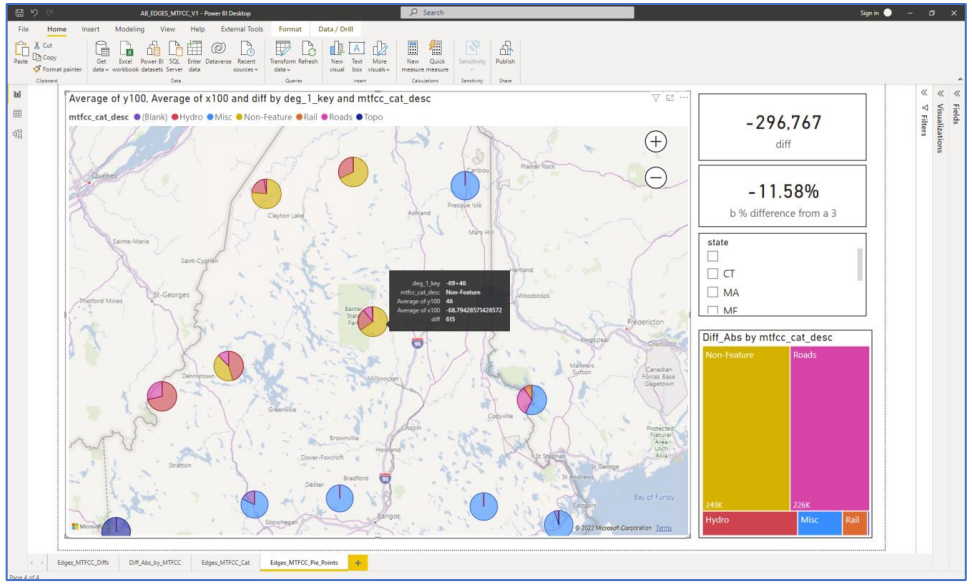
The data change capture and presentation described in this presentation make use of an additional, hierarchical, 2D geospatial dimension to BI aggregation: The Pivot Map, as implemented in MS Power BI.

The use of a set of multi-level hierarchical point grid sets allows for easier, faster, and more detailed analysis, when compared with spatial-join based use of multi-level administrative polygons, which are limited to a single value within any given polygon.

This additional BI based slicing and dicing allows one to start with a high-level summary number: Feature Count, or difference in a number: Feature Count Change, and see it geospatially – show me WHERE the coverage is or where the coverage has changed.

This functionality, including drill-down, roll-up, and drill-through, is generally not available in traditional GIS platforms but can be implemented through customization. GIS practitioners have traditionally not seen data in a BI context.





The End – Thank You - Q & A

Contact Brian Hebert, ScribeKey, via LinkedIn.