Carten Cordell:

Welcome back, everyone. Now let's hear from Dan Keefe, Geography Division's Assistant Division Chief for Address Content Coverage and Geospatial Reference Data at the US Census Bureau. Dan, the floor is yours.

Dan Keefe:

Oh, sorry. I'm going to give a technology presentation and I was on mute. Good morning. Thank you for the invitation to be a part of the GEODATA and GIS workshop. As mentioned, I'm Dan Keefe, the assistant division chief in the geography division at the Census Bureau, responsible for address content coverage and geospatial reference. I'm honored to be here and pleased to share with you the exciting ways my team is enhancing the use of satellite imagery and geospatial analysis at the Census Bureau to maintain our MAF/TIGER system.

For those unfamiliar, the MAF is our master address file. It's a nationwide repository of all the residential addresses in the nation, and TIGER is the topologically integrated geographic encoding and reference system. It's basically a national geodatabase of address points, roads, boundaries, imagery. Before we go on to the next slides, I'm going to go off camera so you don't see me going between screens or yelling at my dogs to be quiet. I'm going to go ahead and go dark here.

Next slide please. This is a snapshot of our MAF/TIGER system. The geography provides the foundation for all data collection, processing, tabulation and dissemination activities for the Census Bureau. It continues to grow just as our nation does, and we must keep up with that growth. This includes new addresses, new roads, new and updated legal and statistical boundaries. As you can see from the slide, we have quite a bit of data with over 14 million unique geographies including legal, administrative, statistical areas. We have approximately 7,000 miles of roads, but also legal boundaries from almost 40,000 governments. We have more than 144.5 million housing units and structure points for approximately 94% of those housing units. Next slide, please.

To have a better understanding of the enhancements and innovation my team is doing that I'm going to talk about today, it helps to know how the census conducted address canvassing over the last three censuses. Traditionally, address canvassing has ensured that the Census Bureau can refine the MAF in advance of a decennial census. Typically, it occurs in years ending an eight or nine. The canvassing is designed to validate the address list through change or adds or delete any addresses as necessary. In 2000, 100% of the infield canvassing was performed through paper-based operations using books or printed address registers and printed maps in the assignment area level of geography. In 2010, 100% of address canvasing was performed using a handheld computer device that captured address updates and GPS and relative coordinates.

In 2020, there was a major shift in this methodology. 100% of the addresses in the address frame were reviewed during an in-office operation. It was called interactive review and maybe some of them have seen it, but this operation resulted in 65% of the addresses being validated as stable, which means that we did not identify any problems in coverage. 35% of the addresses were identified for infield address canvassing. Next slide, please.

The manual change detection that I was talking about in the 2020 in-office address canvassing operation involved manually comparing two sets of imagery using a swipe tool to identify change or coverage issues. We built a block tracking database based on census tabulation blocks to record the results of the review. We conducted a complete review of all the blocks, just over 11 million in the nation, some of which were reviewed multiple times to evaluate ongoing change. We get updates from the United States Postal Service delivery sequence file that we process every six months, so our database is changing, keeping up with growth in certain areas.

In total, we conducted more than 13 million reviews. We identified growth, decline and built out areas. The built out areas were flagged to minimize the need for field canvassing. Saying if something's already built to the maximum in that block, doesn't have to be canvassed. The effort required 150 full-time staff over a four-year period, but the effort reduced the number of housing units requiring field visits by 65% and significantly reduced the number of field staff required for canvasing. As you can see, only 32,000 in 2020 and 150,000 used in 2010.

What is it we want to do? We want to automate the review of the imagery and in-office updates to the MAF/TIGER system, thus increasing the quality coverage and spatial accuracy of MAF/TIGER and increasing efficiency of field operations, optimize resource allocation. Next slide, please.

To accomplish what we want to do, we are focused on the use of new tools and methods, including freely available and regularly updated satellite imagery, the use of parcel data that now includes pretty much nationwide coverage and an incorporation of data science principles and machine learning. Let's get into some of the cool stuff that my team is doing. Next slide, please.

Building on the successful efforts leading to the 2020 census, the geographic support program, or GSP, has already developed several state-of-the-art methods and technologies to identify both where address change is occurring and to ensure that the MAF/TIGER system accurately reflects this change. The Analytics and Change Detection Corps Working Group, yes, ACDC, was born out of a capstone project from the Census Bureau's data. The group is working on automated change detection, parcel matching, and the intelligence database. I was honored to be one of the mentors to the team working on automated change detection, so that was a good learning experience for me. ACDC develops and execute automated models, methods, and implements information storage systems that enable geo to identify change on the landscape such to identify potential address and feature coverage issues that need to be addressed. Next slide, please.

Our primary focus is identifying land that has been cleared for potential construction or construction that has recently occurred. This will allow us to focus on areas with coverage issues and where updates are needed to our MAF/TIGER system. These data, along with other data sources, will help us validate and automate updates. Our first stage of automated change detection is the moderate resolution based on Sentinel-2 imagery, which has a five-day revisit period, multiple bands of data per pixel and manageable size in comparison to high-res imagery, and the observational products for end users for remote sensing analysis, or OPERA, disturbance product, which includes vegetative and general surface disturbance disc products. Where we have persistent change, we will intersect with the grid system to indicate where we need high resolution imagery.

Once we have current high resolution imagery, we'll extract building and roads based on machine learning models and features that are extracted from several images and the most spatially accurate or aggregated into a building footprint layer. These footprints are then attributed using parcel data and other sources. ACDC plans to run an ongoing moderate resolution automated change detection process on the nation benchmarked every six months. That goes along with our regular benchmark cycle for product creation.

We also recently awarded a contract to use Google Earth Engine, or GEE. GEE provides the entire Sentinel-2 library as well as resources to automate change detection models. ACDC will utilize GEE to run a continuous moderate resolution change detection using Sentinel-2 10 meter optical data. GEE can also pull in ancillary data sets that will be used in conjunction with our internally produced change areas. Next slide, please.

We're exploring a couple of methods to identify change with the moderate resolution imagery. One, using a more traditional remote sensing method through the normalized difference of vegetation index. We create a long-term mean average of all available and valid observations on a set of data to describe

the normal state, you can see the normal state is based on a 20-day window, like for example, March 1st to March 20th.

It requires evaluating approximately 10,000 image tiles, creation of the NDVI for each image, create and store a mean NDVI and standard deviation for each 20-day window for each tile. Compare current state to long-term mean NDVI. Collect recent data image tile and create NDVI. Change measured as compared to standard deviation for the long term mean or thresholds. Morphological and spatial filters applied to reduce noise. Change magnitude, direction, and persistence recorded storage raster. Create change regions of interest for further processing and cluster filter change product to create small regions of interest. Sounds simple, right? I asked my team of data scientists for something that I could put on a slide that would be simple enough. They gave me this and they are doing excellent work. Next slide, please.

Back one slide, please. There you go. The other approach that we're exploring for moderate resolution is through machine learning. We are training the machine learning model on reviewed labeled areas from an operation that we have called the New Construction Imaging Review Project. We actually have a team of contractors doing this work now. We compare it against the Sentinel-2 imagery associated to the review. Then we run validation on the machine learning model in different land cover regions across the country to determine if additional regional models are needed to get the appropriate results.

As you imagine, not every model works in a country the size of ours with the varying terrains. We can predict change areas with these change models, the trained models. Collect two vintages of the Sentinel-2 imagery to perform a dual imagery change detection comparison. Then we run the associated model and the two vintages of imagery to produce a change product. It calculates subtle statistics for area of change within each census block. Each vintage of imagery could be a collection of images depending on cloud cover and the need to fill in gaps. We believe the NDVI machine learning and the NASA OPERA data in combination will help us validate and boost our confidence in the modern resolution change that we identify. Each method will have strengths and weaknesses in certain parts of the country or in different types of construction. Next slide, please.

This slide is a visual of the high resolution flow that I mentioned previously and how we get to update. The moderate resolution identifies areas. We request the high-res imagery, run the model, and extract the features. These are then sent to our MAF/TIGER system for update. Next slide, please.

We are also enhancing our use of parcel data. Our primary source of national level parcel data right now is LightBox. We access it through the Homeland Infrastructure Foundation Level Data or the HIFLD program from the Department of Homeland Security, open source data. Other parcel sources are being explored. We're very close to letting a contract to purchase from a vendor, a parcel data set. We take the parcel data and we pre-process it. We look for incomplete addresses. We look for duplicate addresses. There are sometimes stacked parcels where one address has the same geography over and over, think of multi-units.

There needs to be some cleanup done there, but once we clean up the data and validate it, we match it to our MAF, the master address file. This process describes the relationship between the parcel's geometry and addresses to our MAF addresses and our structure points, and allows us to ask and answer some questions like do they match? Are they in the same location, and where are the differences that we need to investigate? We summarize these results at the block level. We talked about blocks, they're consistent. We know the boundaries. We can go back and compare. We summarize at the block level and feed it into the intelligence database for further analysis. Next slide, please.

How are we keeping track of all this work? Well, we stored in the previously mentioned intelligence database, or the ID. In order to better document and assess our work to enhance the MAF/TIGER system, ACDC has developed an intelligence database. The intelligence database is loosely based on the

block tracking database I mentioned earlier from the 2020 work, but this is way more data, way more functionality. This database will be continuously updated throughout the decade with data sources from inside and outside the Census Bureau. The system will integrate the result of change detection and parcel matching, along with statistics from multiple other data sources to help identify, investigate, and track where growth is occurring and where it exists.

The intelligence database will also help us distinguish blocks that do not need updates. This will reduce the chance of introducing errors into blocks and/or housing units that we already consider to be optimal. The intelligence database provides users many views of the data to enable better investigation of issues and trends. Let's take a look at a few examples now. Next slide, please.

The block level intelligence database web application, because we also have it as an Oracle table, allows us to target specific areas where change is detected and the master address file has a potential coverage issue. Here you see a large census block outlined in purple. You click, you see we ran automated change detection in December of 2021, and the process identified growth, the red area there is our area of interest where we identified growth. If you click one more time, so we flagged the area and then we returned using the Sentinel-2 imagery at the moderate resolution. Here you can see the development in this census block is nearly complete. Click one more time, please.

The intelligence database tells us that we have 161 MAF units in this block. As indicated in imagery, this appears to be a major undercount. One more click, please. The USPS's delivery sequence file, which we update our database at least twice a year with, provided or confirmed those 161 addresses, which are geocoded to this block in our system, but we still have an undercount. One more click, please.

When we add the parcel data layer, we noticed there are 846 parcels with addresses in this block. By matching the parcel layer to the MAF, 685 addresses were geocoded and added to this list. Next click, please. The intelligence database indicates that the last field operation to visit this block was non-response follow-up in 2020. If you remember that the automated change detection date was 2021, which is when we could observe the actual change occurring. Next click, please. This block is an example of our automated process in action. The automated change detection identifies an area of change. We used the parcel data in our existing MAF data and 685 un-geocoded addresses are assigned to a block with a structure coordinate derived from the parcel center. Next slide, please.

The dashboard with county aggregated statistics is also available in the intelligence database to better understand a county's general health according to particular metrics. The first dashboard shows us the amount of automated change detection between 2020 and 2021 by actual square meters. Red equals four change. Click, please. You can see Maricopa County, Arizona, and nearly 30 million square meters of change detected, and that's just between 2020 and 2021. Click again.

The next dashboard is based on the delivery sequence file coverage for each county and the nation. The darker blue counties have close to 99% DSF coverage, whereas the darker red counties hover around 50% DS coverage. Click again, please. Users can also zoom to an individual county in these dashboards to view associated intelligence database statistics aggregated for the county. For instance, we see here that one of the counties with the highest DSF percentage, Stafford County Virginia, at over 99% had 726,000 square meters of change detection between 2020 and 2021.

Knowing the quality of our sources in a data rich county such as Stafford will allow automation to handle what was once a staff intensive update process. As you can see, the intelligence database provides a lot of great information that enables the user to drill into specific area to identify trends or examine which areas need to be updated in the MAF/TIGER system. The intelligence database also serves as a fantastic repository of data to help build predictive models to identify what types of data are best at predicting the change that needs the most attention. Next slide, please.

I've demonstrated some ways that the geography division is keeping up with growth, but we also need to know areas impacted by natural and man-made disasters. Unfortunately, we are experiencing more natural disasters than ever before. It is important to understand the impact of these disasters and then to track the reconstruction of the impacted areas moving forward. The intelligence database helps us do that by ingesting and displaying new incoming data sources. This allows us to remain flexible and open to important data produced by other government or commercial agencies. Using post-disaster event data from NOAA and FEMA and National Weather Service and the US Forest Service in comparison with our intelligence database, we can analyze areas impacted. For example, NOAA's damage assessment tool was recently incorporated into the intelligence database to track the impact of recent weather related events.

For example, what you see here is the path of the Rolling Fork tornado that hit Mississippi on Friday, March 24th of this year. This tornado cut across two counties with winds up to 170 miles an hour, and a funnel as large as three quarters of a mile wide. The geographic area in purple are all of our census blocks affected by the tornado. The triangles are survey points along the tornado path from NOAA. As you can see from the summary on the bottom there from the dashboard, 146 census blocks containing 805 MAF units were in the tornado's path. If we zoom to the block level, we can reveal the impact of this event. Click, please.

Here is the tornado path as it rolled through Rolling Fork. Click again, please. I've seen this slide a few times and it's still, every time I see it, it's hard to imagine or believe. This is the post tornado damage image at Rolling Fork. Click, please. With the block level and address level statistics found in the intelligence database, the MAF and incoming source data, the geography division can monitor and update these areas as needed. As Rolling Fork rebuilds, the geography division will be able to track the recovery through sources such as automated change detection output, the information on the delivery sequence file, such as the delivery point field, which indicates whether an address is receiving mail or not. Monitoring all of this data will allow geo to know when homes are rebuilt and mail delivery has resumed for areas impacted by disasters such as this.

Also, during our peak decennial enumeration activities, this type of tracking can alert us to areas where an alternate enumeration strategy should be considered if the impacted area is no longer receiving mail. A quick note, we have done similar tracking for this for hurricanes and the wildfires also, not just tornadoes. We are doing all the disasters that we can get data for.

Next slide, please. Oh, well thank you. I appreciate your time and allowing me to share some of the interesting innovation that we're doing at the Census Bureau.

Carten Cordell:

Thank you, Dan. That was a great presentation. We're going to give a little time for questions, but I was just curious, in the one slide you showed the exponential change of the census as technology tended to grow over the past decades. Not to ask you to look into your crystal ball, but with the capabilities that you have now and the speed of technology, where do you think you'll be in terms of the 2030 census?

Dan Keefe:

Great question. The geography division is writing a recommendation that will be available I believe at the end of next year for public consumption. We are writing a recommendation to propose there would be no infield listing that we would do 100% in office using all of the sources. We all know there's an explosion of geospatial reference and source material online, but we'd also have some partnerships and outreach to get some of the areas without sources. But our recommendation will say there's no need for a large frame building activity to occur in the field for the 2030 census.

Carten Cordell:

Excellent. Why don't give it one more time for questions. Well, thank you very much, Dan. Now I'd like to turn it over to Google Cloud's Aaron Gussman for a preview of his team's work in the GIS space. Aaron, over to you.