UC GIS Week
Tuesday, November 19, 2024
1pm-2pm
GIS & AI: Tree Range Maps, Drones, and Digitizing

Sessions:

All right, everybody. Welcome again to our first of our two afternoon sessions of day one of UCGIS week. This session is on GIS and AI, free range maps, drones and digitizing.

Our first presentation is Sky High Insight, campus facility management with drones by Jingyi Huang and Taran Birk, both of UC San Diego. Feel free to take it away.

Sky High Insights: Campus Facility Management with Drones

Speaker(s): Jingyi Huang and Taran Birk - UC San Diego

Abstract:

This lightning talk will explore the integration of drone technology in university campus facility management. We will discuss how drone flights capture high-resolution images, enabling the creation of detailed 3D models of campus infrastructure. Leveraging advanced image processing and machine learning algorithms, these models facilitate enhanced data analysis for maintenance and planning.

Key benefits include enhanced accuracy in monitoring facility conditions, proactive maintenance strategies, and optimized resource allocation. Through case studies, we will demonstrate how to prepare for drone flights and illustrate how the integration of drones with AI and machine learning can transform facility management, fostering smarter and more sustainable campus environments.

Transcript: Video Timestamp: 0:41

Hi, thank you. I'm going to share my screen. Okay. Good afternoon. Good afternoon, everyone. Thank you for joining us today. We are excited to explore an innovative approach of campus facility management. An approach that brings together cutting edge technology with Rotting

Tasks. Welcome to Sky High Insight, campus facility management with drones. My name is Jingyi Huang. I'm a GIS analyst in the OMCP department. And we have Taryn joining us today to present the project as well. Hi, my name is Taryn and I'm a fourth year college student here at UCSD, as well as an urban studies and planning student. I'm also a student worker in the OMCP department for UCSD, assisting with a variety of GIS related projects. This year I've been working with my colleagues in E to create a 3D model of our campus using LiDAR data collected by drones.

Light detection and ranging, also known as LiDAR, is a form of data collected by a few specific technologies that can gauge points and distances to create high resolution 3D images.

Our tool for collection for this is the DJI Mavic 3 enterprise drone, which is equipped with cutting edge technology for high precision LiDAR data collection. The DJI Mavic 3 enterprises features a four over three CMOS wide camera, ensuring detailed imaging and data accuracy. It has a flight time of up to 45 minutes, allowing extended coverage in a single session. And its advanced obstacle sensing ensures safety during complex math and tasks. These features make an ideal choice for capturing the intricate details of our campus and transforming them into a 3D map.

In order to assure successful and safe piloting, we have both obtained our federal aviation administrative unmanned air system licenses, which certifies us to operate drones for research purposes. This certification ensures we follow all regulations for safe and responsible drone usage, including navigating controlled airspace and minimizing risks to people and property.

After proper preparation for our flight, we start to prepare our site. This includes finding an app location for takeoff that will give us visibility for the entire site plan to map. Then we add ground control points via a Bad Elf GPS receiver, which are markers placed on the ground to enhance the accuracy of our 3D model by providing precise reference points during data processing. With everything in place, we execute the flight. With our first flight this year being at our six college location, we will first take off visible in the video on the slide.

Capturing this high resolution data that allows us to transform the campus into a detailed and accurate 3D map. After the drone flight, we carefully review each image to ensure non-sensitive or inappropriate information has been captured. Once this is done, the images are uploaded to specialized software for further processing. There are different powerful drone image processing tools available, including SiteScan for ArcGIS, ArcGIS Joom 2.0 map, ArcGIS Pro, and MapsMakeEasy. For this project, we primarily use SiteScan for ArcGIS to process the drone data. SiteScan for ArcGIS is a cloud-based drone mapping solution developed by SRE. It allows users to capture, process, and analyze drone data, making it an ideal tool for serving, monitoring, and expanding large-scale projects. Once the drone data is processed, it is transformed into various form maps, including True of the Photo, DSM, DTM, point count, and 3D mesh. Here is an example of the 3D mesh generated by the drone data. The 3D mesh can be integrated into maps, providing an immersive view of the data. As seen in the image on the right, zooming in on the mesh creates a sensation of walking within the scene. However, this also highlights a limitation of the drone flight path we initially planned for the project. The images were captured at around 200 feet above the buildings, which means some flight details, such as balconies, appear blurred. Moving forward, we are refining our flight path to capture more detailed images, while ensuring compliance with UC drone regulations. And currently, drone image data is used to create 3D models for maps and web applications. Moving forward, after conducting additional flights, we plan to explore the use of drone image in conjunction with GeoAI to gain deeper insights. By integrating drone images into AddressPro, we can use the related tools to label the objects, change machine learning models, and deploy these models for future use. As new drone image data is collected, the model can be applied to automatically analyze and detect key features. This process can be used for object detection, such as trees, strict size, and damaged buildings. Additionally, it can be used in predictive tasks like forecasting floats, storm debris, and solar power potentials, as well as pattern detection, such as we can use it for identifying the high injury road networks, 911 call hotspots, and abstracted roads. In conclusion, integrating drone data with GeoAI for campus facility management could enhance the efficiency of monitoring and maintaining campus infrastructure.

Drones' equilibrium sensors and cameras can capture detailed imagery of buildings, utilities, and outdoor spaces. While GeoAI can analyze this data to identify maintenance needs, check the structural layers, and optimize the energy use. This combination allows for monitoring, predictive maintenance, and more effective space utilization. The benefits include reduced operational costs, improved safety, and a more sustainable campus environment. Also, a better informed decision making for long-term planning and resource calculations.

Thank you all for your time and attention. I appreciate the opportunity to share this exciting possibility with you, and I look forward to any questions or discussion you may have at the end.

Thank you so much. And for our folks watching today, please make sure that you post questions in the chat, and we're going to do our Q&A at the end of the session after our three presentations happen. So we'll take our questions after all three presentations happen. Thank you again.

Building accurate range maps for California’s 104 tree species

Speaker(s): Clancy McConnell, Jason Yuen, April Engelmeier, James Thorne, and Ryan Boynton - UC Davis

Abstract:

Biogeography is the study of the distribution of biota across space. In this way, it can inform every kind of ecological study. So, if we’re going to study biota across space, including predicting where those biota might be in the future or how they could respond to external

pressures, we need a good estimate of where those biota are now. In essence, we need a good species range map.

The last major effort to systematically map tree species ranges in California was done by Griffin and Critchfield (1972) of the U.S. Forest Service, over 50 years ago. Since then, thousands of studies have been completed with coordinates for trees, and many thousands more civilians have collected GPS points in public repositories for citizen science, advocacy, or enjoyment. After vetting for accuracy, this data can significantly enhance existing range maps. Also, recent high-quality vegetation mapping efforts by public agencies can provide both additional "presence" data to expand range maps beyond existing known boundaries and "absence" data in the sense that they can be used to subtract out unlikely range area within existing known boundaries.

In this presentation, we'll discuss the partially-automated GIS workflow we've developed for building the most accurate and comprehensive range maps for California's 104 tree species, from creating the highest-quality-ever digitization of the Griffin and Critchfield (1972) range maps to collecting, processing, and incorporating hundreds of additional datasets to subtracting out non-range areas. We'll also discuss how we're using these state-of-the-art maps to assess the climate exposure, fire risk, and operational priority (seed demand) of each tree species to inform seed collection efforts for reforestation, as well as examples of how the actual map-building process can be used to address some fundamental questions in biogeographical theory.

Transcript: Video Timestamp: 9:14

And next up, we have a team from UC Davis presenting on building accurate range maps for California's 104 tree species. This group is led by Clancy McConnell. UC Davis and team, take it away. All right. Hello. I am April Engelmeier. I go to UC Davis, and I will be presenting along with Clancy and Jason today on our project, which has the ultimate goal of building the most accurate range maps out there for California's 104 tree species. Next slide. So just to give an overall workflow of the project, the first step in our project is to digitize the Griffin and Crichfield maps that we are working with. This is the data that we are working with. The next step is to collect and clean other point and polygon data sets and make considerations for our range map building. We then plan to make subtractions and additions of data to each different species depending on what we see fit. And lastly, we plan to finalize the range map. Next slide. We are currently on steps two and three where we've begun to collect and clean other point and polygon data sets and make considerations for our range map building. And we have successfully completed the first step, which although it is only the first step, it is definitely extensive. And that is the bulk of what we'll be talking about today. So just to reference, we are going over 106 California tree species, but the Griffin and Crichfield maps that we are working with are only 86. And there are another 15 to 18 species that are also on our radar. And for the sake of both funding and timing for this project, we are prioritizing around 20 species that you can see on the left here, just to make the most effective references that we possibly can in the time that we have.

Next slide. So to go further into depth on the first step of our project, there were three main levels to our project. And the first was scanning the paper tangible maps that we had that were the Griffin and Crichfield maps. Once we scanned them, we were able to geo reference our maps into the data that we wanted. And then finally, we're able to digitize them. And I'll go further into detail on each of those steps. So we started with the scanning process. The original maps that we were provided had a DPI of 300, which is relatively low. We wanted to rescan our maps so that we could both geo reference and digitize with more precision and accuracy. So I took all paper maps to a flatbed scanner and rescan them at a DPI of 600. So that our TIFF files had more precision in our next step, which was geo referencing. So then we moved into geo referencing all of our 86 Griffin and Crichfield maps. The original data that we were working in was NAD 1927, California till Albers, which makes sense because this data that we were working with is relatively old. And we ultimately wanted to project our new data into 1983. So to start with geo referencing,

I would add a TIFF file onto our base map, which had a California state border and gradcules. And essentially using the geo referencing tool, I would line the gradcules of the base map to the crosshairs of the Griffin and Crichfield maps. And this photo on the right kind of gives an example of that. You can see that there are little black crosshairs on the Griffin and Crichfield maps. And we essentially wanted to line them up with the base maps that we could project them. Now the distribution of the maps, obviously for each species ranges, some species had one singular point on the map or one singular polygon. And for those that didn't have a range beyond more than 10 crosshairs, some only ranging between around three and four crosshairs of a map, we just used inset maps. But the majority of our data, we were able to line at least 10 points with the crosshairs on the map. And from there, we're able to use the spline transformation to geo rectify our image. And once the geo rectify, once the image was geo rectified, we were able to then project our maps into a California Teal Alberts 1983 file. And from then on, we were working in the 1983 datum for the duration of the digitizing step. So digitizing was the next step and Jason will take that away. Yeah, so from the initial geo referencing step, we then began this project statement or this project segment, starting with geo reference map displaying the known ranges, stands within a one mile radius and the estimated ranges. And then the finalized map basically came with each level of location categorized into blobs, points and polygons created with dash lines alongside the future classes created during the process. So all we had at the end was basically just a big conglomerate of data. And yeah, next slide, please. Yeah, so our process involves creating different levels to present either home ranges or known home ranges, which were our blobs, stands of species within a one mile radius, which were our points and estimated home ranges, which were our dash lines in the geo reference map. And we used a Jupyter notebook in ArcGIS Pro to help with this process that had all of the geo processing functions needed in the form of Python code that could be called by the user in a very simple manner, which you can sort of see right there. Next slide. So this was the main part that me and April used primarily. The first box allowed the user to input in the species name. And if the species had either a point, a polygon or a dash line, this would be very simple for the user to just be like, you know, you could put in the abbreviation. So in this case, kudo, if it had points, he would say yes, if it had polygons, you can say either yes or no. And if it had those dash lines, you could say yes or no again. And then all the user would need to do from that step is they would need to input in the function title found in that little title above, and then they could start digitizing or, you know, doing the steps for digitizing. Next slide. And then the final result would be something like this, where you can sort of believe this one is just one feature class that we symbolize differently based on unique values. For the points, it would have its own little number for the polygons, it would have its own number and for the dash lines, it would have its own number. I believe it's number three. Next slide.

Yeah. Great. Thanks. Thanks, Jason and April. So that's like the bulk of the work so far on the range map building. And I'll actually go back to that slide. It really deserves to go. We need to go back to that. It was so much work to do this. Just want to emphasize that these two put in hundreds of hours of work, scanning, georeferencing, digitizing this, and anyone who's done digitizing knows it's a huge amount of work and time. It's tedious. You get, you know, eye cramps, back cramps, neck cramps, finger cramps, I mean, everything from just doing it on and on. And, you know, the workflow that they came up with and the script made it so much more efficient. So anyway, this is like most of the work involved with it. These two did. And so it's really impressive and really grateful they came onto this big research project. I'm just going to present the next part, which I'm working on, April and I are working on now, and hopefully we can get Jason back on to help us with some of it. And that is, we now need to decide how we're going to build these range maps. The postdoc that I took over from came up with a pretty good workflow about taking these base maps, Griffin and Crichfield base maps, in 1972, and then adding in new data, subtracting stuff out, sort of the workflow that April was mentioning at the beginning. But before we get to that point, we need to discuss the differences in how we would go about making a range map for different species. And for the biogeographers out there, this is kind of the crux of what you'd probably deal with when you're talking about range building. You know, common species, we have a lot of good public data for that, but they're not really well cataloged in places like Calflora, for example, which is sort of that weird phenomenon, you know, it's common species, but no one wants to put it in a public database, because there's so much of it, it's not that interesting, right? Rare species, sort of the opposite, you know, we don't have a lot of good public data on them, things like the National Park Service mapping inventory. But in Calflora, there's quite a few, especially considering the small range. So that's one consideration is where we're getting data from. And then another one is how we might actually go about the decisions that we make in, you know, the process of adding and subtracting out different data types in the raster building process. And I'll go over those three, biogeographical rarity and elevational distribution differences. So just to hop right in the biogeographical differences, as an example, might be the difference between a matrix distribution where it's relatively continuous, this is Blue Oak on the left, relatively continuous, sort of that nice bathtub ring around the Central Valley. You don't really see it in individual populations. For the most part, it's somewhat continuous around that elevation zone and across the entire landscape. On the right, this is California Buckeye. It's almost the complete opposite. It follows, you can see, almost a very similar distribution to Blue Oak as far as where it's found across the state, almost a bathtub ring it has on its own, but it's not continuous. It's found in metapopulations. It will be on one south facing slope and then you go, you know, one canyon north or one canyon south, and then it's not there at all. There's not a single individual. And so that can impact how we put the range map together. Another one might be the commonality. Is it found in wide distribution in very large populations like Coast Live Oak on the left, where we have a lot of good data for it. We can be pretty confident in what we know about its distribution. On the other hand, rare species, these are only found in a few data sets. They're weird, quirky, eccentric individuals that I'm having to track down for data. You know, there's like, there's a Colter Pine guy, you know, and he's like, he lives in this one city in the one part of the state. No one's seen him in 25 years, but send him an email, you know, that kind of thing. And then another difference might be elevationally on the left. The Griffin and Critchfield maps for our, you know,

our endemic Juglands were really not that good. This is Juglands Hinziye Black Walnut on the top, and Juglands, California, Southern California Black Walnut on the bottom. But they're limited elevationally, just found between 30 and 900 meters. That's pretty narrow distribution. That's really nice. And it's also limited in that we know it's never really going to be found above 900. On the other hand, on the right, we have Jeffrey Pine. It's got a really wide elevational distribution that varies depending on what part of the state it's found in. You know, up in the Klamath, it's found at much lower elevation than in the Sierra. And so we can't just necessarily have a blanket, a blanket limit on where we cut out and add on to Jeffrey Pine. We have to be specific in which region we're talking about. So once we've, you know, had those considerations, then we move to data collection, or all of our additional data beyond the Griffin and Critchfield maps. And this is point and polygons. And just as an example, I've already done this for Black Oak, Corkus, Kalagi. And out of the preliminary data set of over a million points that we've gathered for this whole project so far, we have over 22,000 additional points just for Black Oak. It's amazing. It's a lot of data. So some of these species, there's a lot of good data out there. For the whole project, we've collected almost 100 data sets, processed most of them. Those are point data sets. Point data sets, 92, I didn't put that on there. 25 polygon data sets. So well over 100 data sets for the project. And we're getting these California Fish and Wildlife surveys this year and next year. This is data that's not yet been made public, but we're getting the preliminary version because, you know, our project is due next July. And then there's some other data we're getting from individuals, like I was mentioning. You know, there's that Colter Pine guy that no one's seen in a while. And there's professors and scientists I'm tracking down all over the West Coast. Now from there, we move into deciding what we're going to subtract out. You know, we've got this perhaps nice continuous distribution shown from the Griffin and Critchfield map, but we know for a fact it's not found in a continuous surface in that area. So how do you subtract out, let's say like the riparian zones from that, where we know the tree isn't found? How do you subtract out areas where it's just bare granite in the Sierra, where you know there aren't any trees? We add more data in and then finally do an extraction. This is an example of what it might look like when we subtract things out. So the two main criteria right now are subtracting out range listed in the Jepson manual that's not included in the species known range. So this might be everything above and below a certain point. And then we could also subtract out CWHR types that don't mention that species as a primary species in its description. Now that second one is kind of sticky because, you know, CWHR types are not going to mention a lot of rare species. So it's not a useful, it's not necessarily a useful step to do for the rarer ones, but for the more common ones, a lot of our conifers, a lot of our oaks, that's a useful step.

One, now we get to the adding data back in and one thing that we wanted to make sure that we did when we were adding data back in is somewhat similar to the system that Jason developed in the Python script was we wanted to make this replicable and transparent so that if you were to click on a raster cell, you would be able to know where the data in that raster cell came from. If you were to get the final range map, it would be attributed with a code that could then take you back to a table. And of course, this could be table joined too, but for simplicity's sake, to keep the file small, it could take you back to a table where that has all of the information from the original source file, who collected it, what data set it was part of, the date that it was collected, what was the accuracy, all of that information can then be retained. So if you were to click on a cell, it might simply say this is part of, you know, it might say Griffin and Crichfield. So we know that that's where it came from. Another cell might say, well, this is from the Calflora database. This is who collected that data point in 2007, etc. And that the data could also be symbolized, you know, in different ways, depending on what attribute you use in the final raster output. Then we merge that data back into our source Griffin and Crichfield raster that's been, you know, subtracted out, we add data back in, might be a little bit hard to see on this screen, the Griffin and Crichfield is the gray in the middle, and then the red raster cells are the data that's been then added back in. As a final step, it's not included on this page, but we extract the water bodies, lakes, streams, that kind of thing, anything that sort of might have spilled into the raster. And then we have our final range map there on the right.

So that's kind of the cap on what we've done for the data collection of this project. And as a last short little bit, I just want to show you what we're going to use this data for. And you can come to my presentation later this week as well. Testing the script. This refers to a separate r script that we're using. This is for Coast Redwood. We have an r script that takes these highly accurate species ranges and models their climate exposure. So we take a sample of 100,000 cells, we model their historic climate distribution, and then use three different climate futures, one of which is actually the present. Three different climate futures and model how the climate in those cells is going to change. So it's sort of a place-based climate exposure analysis. We're not predicting where the species is going to go. We're just going to predict how exposed it is here. And we can model this over time. So this is those four different periods in the existing Coast

Redwood range. And I think that's where I should probably end for now. Thank you so much, Clancy, Jason, and April.

We're now going to move on to our final presentation of this section by Erin Mutch of UC Merced. And he's going to present on hand-drawn historic maps, utilization and conversion of unique features into GIS.

Hand-drawn historic maps: Utilization and conversion of unique features into Geographic Information Systems (GIS)

Speaker: Erin Mutch - UC Merced

Abstract:

The GIS Center at the University of California, Merced provides research services to faculty and is responsible for managing and maintaining a geospatial collection for our students and researchers. We support Geographic Information Systems (GIS) software for over 500 users yearly and provide workshops and consultations. GIS Center technical research staff have provided consulting services to researchers and direct project support. Through four case studies, we will detail our processes for converting historic maps into GIS products that contribute to research publications and data sets. Specifically, we will discuss the technical demands and research benefits of georeferencing and vectorizing digitized historic maps utilizing various methods including machine learning and heads up digitizing and sharing that data via ArcGIS Online and other cloud-based applications. We conclude with recommendations for others embarking upon this type of work.

Transcript: Video Timestamp: 26:55

Unmute. There we go. 50 button moved. All right. Well, happy GIS Day, everybody.

Yeah. So I noticed this was about AI and digitizing. I was like, oh, okay. At first I was like, okay, we're on the humanities side. I'm going to share four case studies of work that our GIS Center has supported or led over the past few years. Again, my name is Erin Mutch. I'm with UC Merced. And I'm presenting. This is a presentation based on a recent journal article that was published in Taylor & Francis. Hopefully this will be in eScholarships soon. So I do have the DOI there linked. If you actually want to read the paper, or you could just listen to me for 15 minutes and talk about this cool project or these cool projects. So we started as solely research support at UC Merced. And our goal was to, again, if someone came in with a grant, they'd pay us to do the work and we would try to figure out how to do the work. And when you're embarking on these type of new ventures, you kind of have to take whatever you're given as far as grant funding is concerned and you take a lot of risks. So as you listen to these four projects and I talk about that they were risky, they took a lot of effort. But I do believe that the output for the scholarly community is going to be very beneficial. And I believe there's going to be an increased demand in the use of GIS and digital humanities projects. So in 2016, we migrated to the library because they realized we need a lot of education and training and support that wasn't happening on campus. So I'm kind of a hybrid operation where we are part of the library, but we also do projects. So that's just a little bit about that. And so really, again, the overview is that we're finding a great increased demand that historic maps are very valuable. So our role is to support the process of georeferencing these historic maps and then figuring out how to vectorize these data into GIS features. With the advent of AI and all these tools, everyone's like, oh, you just kind of like do, do, do, do, do, and it makes GIS. It's just magic. I mean, I think we all know that's really not the case. So hopefully what I'll share with this group today is like, how can you communicate like the benefits of using this type of technology versus the constraints and the expectations that a faculty member or researcher may have? So in addition, a lot of digitized scanned maps are becoming more and more valuable in the research community. So I hope that in general, universities will keep investing in map libraries and continuing the work that everyone's been doing there. So you probably all realize or you might understand the georeferencing process or maybe not, if you're new to GIS. But when you work with historic maps, the process of georeferencing is really more of an art than a science. So you can't just run some numbers and algorithms and it's going to be perfect. I mean, it's a bigger challenge with historic maps because a lot of the digital representations of the maps have fold increases. They were hand drawn, they weren't surveyed, they weren't taken from aerial photography or satellites. So it's, they're going to have a lot of unique challenges about them. And then figuring out how to use which choice of projection is also could be detective work. Sometimes these maps were drawn into projection you may not be aware of, or you may look at a map of South America and realize, oh, I should use this projection or that. So again, for those who have studied GIS, worked with GIS, we haven't really talked a lot about projections as much as we used to 15, 20 years ago, because the software kind of auto detects these things for us. But it can be a lot of detective work to get it right. So after we go through a process of data referencing, these are, these data and these maps can now be utilized to create vector features. So I'm going to highlight four case studies. Again, I'm going to go through these probably faster than I want to. And I'm open for questions. You can contact me directly to if you have any specifics on any of these. These, all four of these projects were born out of different departments. So the Mayan Caves were from our anthropology professor. Yellow River was for, from one of our faculty,

sorry, our historic former faculty history and history. The British colonial India project was born out of political science researchers and faculty from multiple universities. And then Cleopatra actually came out of an economics faculty member. And that's kind of the one that is just a lot of work. And it's spanning a lot of different areas, but we're excited about it because it's going to be published hopefully soon. So our first project with the Mayan,

the Mayan ritual maps, ritual maps, these came out of digitized, we digitized features for over 150 cave maps in Belize and compiled data on over 11,000 artifacts. It's a lot of data.

And these maps were created for graphics and map labs for publication. So it was, we went in, we took inventory of this cave and we mapped it and used in many cases analytical tools at kernel density. They're looking into like acoustical analysis in the future. And one single cave could have 50 to 100 shape files. So I'm struggling with my zoom window, I apologize.

And so it's really complex. And when I got this project at first, I was like, well, why are we using GIS for this? This is for graphics. Like I said before, we're studying analytics within these caves and how the caves are being used or how they were being used.

So the biggest challenge of geo-referencing cave maps is they're subterranean. So you can't just get a GPS point and get those geo-referenced. So we don't have any surface features to do any control points. And so what a lot of the researchers did is they took a GPS point at the entrance of the cave and then we kind of just rubber sheeted the maps to match that GPS points. The interior caves were mapped using multiple sheets by multiple researchers. So they would go in with their maps in the cave and they would get dirty and they would get dusty and they would get water along, but we would take those maps and try to figure out how to geo-rectage those with the main outline of the cave that was mapped. And so cave features are really complicated and have intricate database schemas in addition to all these other things we're trying to do. It's like, you guys are crazy. Why are you using GIS? This is why we're using GIS. We're using GIS because now with the data mapped within these caves, you can do these analyses. You can see where there's wear and tear, where there's been impact, or where there's density of certain ceramics, and then analyze potential activities within these caves. I'm not going to go into the details because I'm not a subject matter expert on the caves, but it's still ongoing. We continue to start building, we continue to build the databases to support these analyses. A second more recent project that came our way was the creation of polygons that represented very detailed district boundaries from the decennial India census starting from 1901 and ending in 1931. There were maps that were on Library of Congress, other archival sites, and they said, "Okay, this is really great data because now we have the back-end census data that ties to the sub-districts and we can analyze literacy rates for genders." That was one of the things they were looking at. So where were the investments in education that were just for males or for males and females back in those days? Actually, what they're looking at is how this data has affected real world situations in India.

But again, I'm not the subject matter expert. I'm just going to deliver the data. The final feature classes included an estimate of the error in the metadata. So this was one thing that I encountered like, "Okay, when you put it in a GIS and you distribute the data, people don't know how accurate it is," or they assume that it's more accurate than it could be. So metadata, of course, is really important. And how do you really measure that when you're georecerencing maps? So that was also a challenge to us. And we also coordinated with multiple researchers. So we published not only just the data, but the raster images on a private ArcGIS online group. So we were able to look at the data together and they were able to see what we digitized versus what they're trying to look at. And they were very... I discovered the math scroll, finally.

The disabilities were really detailed. Again, the data, the backend data was really rich and they can use that for analysis for really more in the political science area. So the biggest challenge was these maps from different senses, as we know even in the US census, that these census boundaries change. So one district from 1901 may not be the same shape as 1931. So that's clearly a challenge when those change. And there was an initial expectation that machine learning was just going to automate everything and say, "Hey, we're going to just spit it into the computer and just draw the boundaries." And I did run some automated vectorization on it and ran into some errors. So we did a hybrid approach of machine learning and heads up digitizing of cleaning up these features. And the margin of error was significantly more than contemporary maps we made sure to put in the metadata. And another part that was challenging is that the district boundaries were in Indian language and they were very difficult to read. So actually we had a great student who had a little bit of background in the language with better eyesight than us be able to kind of decipher some of these names and do the research on that. So the output was just this basic polygon boundary that we delivered to the researchers and now they're using that to do their own research. I'm going through these quickly because I know time is limited.

Project number three, this was one of our flagship projects from one of our founding faculty members, was taking yellow river maps and digitizing those to do, again, contemporary analysis of maybe real world conditions or changes of patterns over time. So we took large format maps for the historical atlas of China and they were scanned at high resolution. And the data entry points for these maps included towns, prefecture, fortifications, and garrisons. So this was where

certain tribes with certain populations would migrate and establish themselves. And then

looking at that over the watershed boundary, hills and valleys and all that good stuff, the data created from the historic maps were then utilized and refined to support research publications. So again, we did the data creation and then passed that data off to the researcher and then they took it to the next level. So this was almost 10 years ago when we finished this project. The biggest challenge were when I saw these maps, I thought we can integrate them into a seamless integrated mosaic. Again, these are historic maps. Why was I thinking that was even possible? But we were hoping to achieve that, but we just couldn't get them to match perfectly.

A lot of these maps, as I mentioned earlier with other projects, they have creases and folds and warped. So when you're georeferencing these areas that have these creases and warps or distortions, you either have to fine tune those, but if you fine tune those, that might disturb the whole georeferencing process. So again, that's why I said it's more of an art than a science. And at the time we used ArcMap because we didn't have ArcPro back then, and we had to configure our keyboards for entry of Chinese language characters and also have a graduate student who can read Chinese. So this is one of the examples of the maps that were digitized. Again, I'll try to zoom in. I'll probably blow up this presentation, but yeah, this was high resolution at the time and it's still kind of hard to read. And again, you can see with the folds and the creases, we had the gradicules that was very helpful for the georeferencing process, but obviously there was distortion that we had to just acknowledge and just accept that things weren't going to be as accurate as maybe we hoped. So my last project that is, I think wrapping up, I hope, is this project we've been working on for almost 10 years to create contiguous geospatial historic database from 500 animated maps. And luckily, georeferencing these 500 animated maps, 500 different years spanning 3400 BC to present day. This was a popular YouTube video that had over 5 million views and an economics researcher obtained them and gave them to me and said, "Hey, Erin, can you figure this out?" And I was like, "Sure, why not? Let me try to figure this out." So luckily, the georeferencing images were not the challenge because they all had the same projection they had, the same resolution. But these were basically someone using Microsoft paint to color in changes of political boundaries over time. And our first pass that we started on, I would say 2015 to 2019 is my guess. So it took us four years to do the initial digitizing based on these images. So what happened was, and this is my like, "Okay, when you do this type of work,

when you share the data with researchers, if you don't want it getting out to the general community, you have to figure out a way to tell them to say, "Hey, don't share this without metadata or if you're going to share it, this is not like publishable data." And it really wasn't, but it was like, "Oh, no, we're just using it on the back end for statistical analysis." Well, it was shared out. And so a group of new researchers found it and said, "We want you to make this data better." And I'm like, "Sure, why not? Let's make it better. I can handle that." So a lot of the areas were work leading to random inaccuracies. A lot of the names that were coded were from the legend, needed some subject matter expertise. So the initial data that we received again from this YouTube video needed some subject matter expertise. And the outcome of this, which was probably like, it's probably a two-year process, but it's probably 18 months, a newly updated geodatabase of 165,000 or 100, I don't know, almost 200,000 polygons. But the challenge is now it's too complex for basic visualization. So we're working on that. So this was a panel from the video showing changes over time. So I don't know what year this was. It's probably the top corner. But this was the data that we took, we automated and we digitized. And the original digitized polygons looked like this. So our job was just to match this image. We weren't trying to do anything else because we weren't given ample budget or time to do anything more than that. So this was our original digitized. Then as I was aligning, this was being the next panel, this was the machine learning digitized polygons. So there was machine learning code utilized to draw these polygons out. Again, you can see there still is a shift and there's some questionable things happening. This is the data that we're hoping will be released by the end of the year. This is closely aligned polygons that were ground truth with subject matter expertise, showing the details of historical changes over time. And I know I'm hacking up the details and the subject matter expertise of all four of these things because I'm trying to rush through them really quickly. But this is data that is going to be open to the scholarly community. We submitted this to scientific data nature journal. So we're in revision of our paper. Hopefully, they'll approve the paper and that will be released very soon. This is going to be open source and available to everyone.

And to the left is an example of the alignment process that I had to go through. I tried to automate it. I did automate it here and there. But I used ArcGIS Pro to go through and check these boundaries out. There was a lot of hiccups with ArcGIS Pro crashing, too much data. So there was just a lot of nuance and tricks up my sleeve to figure out how to get through the alignment process. So the conclusions of all of what we worked on over time, and again, I apologize if I glossed over things. Because we just finished this a couple of weeks ago, I'm like crashing and burning really quickly, as I'm sure many of us do after a big project. But historic maps can be converted into data. So it's possible. And there is a demand for this. I believe it's valuable. But in digital humanities, there's not a lot of funding. So you've got to take these projects on either as a labor of love, or you have to go in and take some risks. If it is well funded, then good on you. I'm happy for you. But a lot of the stuff that I stepped into was very risky. And I took a lot of risks to kind of make some of these things happen. And I'm hopefully with that risk, there will be reward to somebody in the future. I don't know who, but hopefully, if we get our project published, that would be nice. Distribution of GIS, again, document your data and make sure that when you share it, put the proper attribution for when used for research. So just, and again, I don't have a playbook for this yet. I'm going to work on that because I've already been through it. And I know there's resources and support out there. But I think with GIS data there, again, there's this misunderstanding that just because it's in GIS, it's good. Or back in the day, people used to think, oh, GIS is like a cartoon. It's not survey data. Well, it's garbage in garbage out. Like your GIS can handle as accurate data as it's going to put in or the computer can handle. But just be aware of those things. Georeferencing is not one size fits all. And I think, you know, map librarians, GIS analysts, humanities researchers could really leverage these resources for innovative research. So thank you for listening to me ramble about my four projects. Feel free to reach out to me. I'm going to check the chat for any questions that you might have for me. And thank you. Happy GIS Day, everybody.

Thank you so much, Erin. I'm particularly fascinated by Georeferencing caves. I can't even, that seems extremely challenging. Yeah, I try not to get, yeah, it's not easy, but

Q&A

we're making it work. So we're having some great conversation in the chat already as some folks are asking questions and interacting with the various presenters. We do have a question from Hugh Livingston for the UCSD team asks, can these models be created with desktop computing power or is it much more demanding? And as I say that Jingyi pops in with a response that, yes, ArcGIS drone to map advance could be the desktop application available for processing drone imagery into 3D. Do other folks have questions that otherwise haven't been addressed?

I also want to remind folks that we have a Slack channel for UCGIS Week 2024. So if you're someone like me who as soon as a Zoom question ends, immediately has questions pop to mind, feel free to share some of your questions there as well. We can hopefully have some good discussion over there throughout the week.

Yes, a big thank you to all of these excellent presenters in the time slot. Big virtual round of applause for all these folks. Really, really fascinating. And if there are no other questions, we will be back here at 3pm for a series of presentations on the urban environment from Latin America to California. So I hope we see many, if not all of you back here in a little over an hour.