Northwest Anthropological Conference
28 March 2013
Transcript of Braden Elliott's talk

Good morning, everybody. Thanks for the kind introduction, Alex. My name is Braden Elliott, I'm an MA student at Oregon State University, and as I walk through the valley in the shadow of thesis, I want to share with you where I'm at right now. This is a very map-driven presentation, so here is a map of where we're going today. I'm going to state a lot of obvious things, to many of you, because I'm sure there is someone in the room that doesn't know every obvious or redundant thing that I say, so humor me for doing some things at a very basic level—I apologize for that.

What is camas?

We're going to start off with the subject of what I'm looking at, which is camas. Camas is a flower, many of you might know it as a blue flower. It blooms in the late spring and early summer, lower elevations first, and moves up the mountains as it goes. The background (of the slideshow)—if you forget what camas looks like, you will continue to see it in the background for a quick refresher. Camas is a pretty important plant, thinking about the prehistory of the Pacific Northwest. We find it in the Pacific Northwest, but not really outside of the Pacific Northwest, insofar as you find the edible version. You find different species within the genus *Camassia* other places, but the bulbs were not traditionally eaten as a food source.

You would dig up the bulbs, cook and steam them for a couple days in an earth oven. You could eat them immediately, you could put them into cakes and dry them. These would be stable for, potentially, years. It's a trade item and I've heard it compared to salmon in terms of salmon being the most important animal food in the pre-contact Pacific Northwest. Camas would, similarly, be at the top of the pyramid...or, I guess, the bottom of the pyramid, if you think about it like the USDA food pyramid, for plant foods in the Northwest. I think that's all you really need to know. I could get into the genetics a little bit—that's what got me interested in camas in the first place—but it's more detail than you probably want because it's not exactly related to what I'm doing right now.

Measuring indigenous natural resource management

We're back to the map of where I'm going. Most of the presentations; well, the two presentations so far, I've seen a trend that there's a field season involved and you get the whole rainbow of "we started here, we did our work, we ended here, and now here's our analysis of the thing we did in this bounded timespan." You only get half the rainbow with my talk—I'm sorry, I'm not done with analysis yet because computer programs don't have a field season and I can keep working right up until I defend next month. This is more of a status update.

There have been a couple attempts in the past to measure indigenous management of camas as a natural resource. I would like to point out two. Stephenie Kramer wrote her Masters thesis at University of Oregon, defended and submitted it in 2000, and she was looking for increase in the size of bulbs (excavated archaeologically from ovens in the state of Oregon). She did not find significant results, and I would like to point out that there were also a number of things that were not controlled for, that, in most cases, I don't think could be controlled for. For instance, what's the ecological setting around an oven? What are the predominant species around an oven? What's the management strategy around an oven? You can't be on the ground to know those things, and it's very hard to infer them archaeologically—especially when all you've got are charred remains that happened to survive thousands of years to be excavated. It's very possible that noise in the data could totally undercut that before the researcher had a chance to do positive things or make mistakes.

On a separate track, there have been a couple genetic investigations—the ones that pulled me into this in the first place. Susan Kephart, who is a biologist at Willamette University, has been involved in those research teams, she just doesn't show up as the first author on most when you look at the citations. Since 2006 she has continuously been part of teams that are sequencing the DNA (mostly from chloroplasts) of vegetative material off the leaves of plants, to see if there is homogenization within human boundaries and differentiation between human groups on the landscape. In my opinion, the human component of those studies has been a little underdeveloped. I'm willing to write that off as the studies having been conducted by botanists—people trained as botanists—so I'm not going to tell them how to do molecular botany, but they're not doing a very good job on the anthropology side. I have no experience with their contact with indigenous stakeholders to see if they're checking their methodology, either. The picture there is pulled from the article Hiroshi Tomimatsu first-authored in 2009. They were studying the Puget Trough area, looking for camas to be more similar genetically within human groups than between human groups. They were using death camas as a control, which I think is an interesting point, because death camas might not have been eaten—but I would argue that it was certainly managed. Weeding something out of your fields is management; and there are medicinal uses like as a topical for cuts, burns, and stings; you could use it as a regurgitant. Management can take different forms, so if you're trying to compare one against another—in terms of whether management is present—without differentiating the kind of management, that might also contribute to why this study also failed to find significant results.

These two methodologies have been been thrown at the question of "how can we measure the impact that humans had on camas, before Euroamerican contact, in something that's visible today?" How can we find a lens to look back at that, and then quantify it in some way that's meaningful—in some way that we can, say, do math or make statistics out of it?

Habitat Distribution Modeling, a brief introduction

I'd like to introduce Habitat Distribution Modeling to you, to any of you who have not had any experience it; then we're going to go through how it applies to the question of pre-contact indigenous management of camas.

My current favorite method of Habitat Distribution Modeling is Maxent, which stands for "maximum entropy," which is a really fancy way of saying the computer knows nothing at first. It assumes that everything is chaotic and there's a uniform probability of anything happening anywhere. You have to coax it away from that state to talk about what the range of possibility in the real world is, and then you have to pull it farther with what you know about where things are on that landscape (which you've now convinced the computer is there) to talk about why they're those places on the landscape.

Let's break that down a little bit. The end product of Maxent is a habitat suitability map. Most of these maps (in the presentation) are pretty basic, I don't have all the basic elements like a scale bar and North arrow—hopefully, most of you know what the state of Oregon looks like, and can interpret these maps as such. This is a pretty simple color ramp: green means it's suitable, red means it's unsuitable. The way that's created is by scoring each pixel based on every variable that we know for that pixel, based on the way camas is observed to behave in this total area of interest.

We need two basic sets of data. We need a bunch of dots for where camas has been found, and then we need a bunch of layers of what the environment has been measured as. You're probably already noticing a lot of generalization, or inference, or noise in the data, which I am having to deal with as well; we're going to see if I can avoid the pitfall of statistical insignificance in the end.

What do we do with it? We start with a background sample: the computer pulls a bunch of points out of everything, measures all the variables at all those points, and says "here's the characteristic picture of what the environment is like." Elevation doesn't go from negative infinity to positive infinity, it goes from something like negative five to thirteen hundred arbitrary units or something. Now we have a range, we have a distribution (or density) of those things in those ranges, and the computer has gone from "anything could happen anywhere" to "this is what the landscape looks like.” Then we pull data from the points where we know a species has been found at some point, and make a similar curve. We now know if it spikes at a particular value on any variable, and if that spike is in a different place (or at a different magnitude, but mostly at a different place) on the histogram than a spike from that same variable on the background sample, we know that we find it in a different place from the total range of variability, of from where it's most likely to be found in if we assume random probability. The more different those things are, the heavier we make that variable in the model. The less different those things are, the lighter (or less consequential) we make that variable in the model.

Then, we move forward: we look at every pixel, take the value on each variable, weight them accordingly, turn each value into a suitability on that variable, then lump them all together and say “altogether, this pixel is probably 80% suitable—to have habitat that camas would like.”

That picture (on the slideshow) has some detail you can't quite see. The are three dimensions there. On the floor of the diagram are the two measured variables, so it's sort of like a three-dimensional histogram. The height is just how many points you find at that intersection of variable scores (or values). This image came out of an article that Janet Elith put out in 2011, which is called a statistical explanation of Maxent for ecologists, and it's pretty dense—but it's kind of fun to read if you like, I don't know, reading Wikipedia on statistical methods (which is one of my new hobbies.)

Modeling camas habitat

So, how do we apply that to the question of human management of this plant? We need to get our data sources. You saw them in the previous slide, you're going to see them again here: those little green dots latitude and longitude realized on a satellite image of Oregon. Those come from the Oregon Flora Project, which is housed at the herbarium at Cordley Hall at Oregon State University. So, I'm using some locally-housed data sources, but they're available through the internet—you can download them directly from oregonflora.org/altas.php, get a spreadsheet, cut out all the information you don't need (like the “taxonauth”—unless you're a botanist going to do you much good for this process). Now, you have spots where camas has been found...sometime between 1880 and the present. We weren't trucking around with a Garmin GPS, or a Trimble, in 1880; some of those points can have up to a 12 kilometer margin of error around them because (the reports roughly read) “we were in this canyon when we found it, and we were sort of on this road...” so we just put a point in the most likely spot and say that somewhere within this 12km radius is the spot.

Here is the first quantifiable uncertainty in the model. We could take a sample of those points, or all of them if you really want to crunch some numbers, and talk about how much spatial give there is that you need to account for in the model. Quantifying the uncertainty is an important part of modeling. All models lie, some models are useful—insofar as you can apply what they're telling you appropriately, and only apply them as far as it's actually safe to go.

Now we move to the environmental variables. The PRISM Climate Group, also housed at Oregon State University, pulls data from weather stations. They do this monthly and put their results online. They interpolate a surface based on those values (and topography), so we can there get another error measurement: How does their interpolated surface compare to their measured values? We can have, say, a standard deviation of error there. Next we take elevation, from the National Elevation Dataset in the form of a Digital Elevation Model (DEM). From there we can use tools such as the handy toolbox in ArcGIS (to generate slope and aspect).

I guess I never stated this obvious one. GIS stands for Geographic Information System, which just means a piece of technology that handles spatial data, and a person interacting with it (at various levels of expertise, and the system working with various levels of bugginess). It's a lot of fun, it's an adventure if you're into that sort of thing. You can pay a lot of money for it, you can do the open-source thing and read these really arcane online help forums, it's your choice.

You can derive slope and aspect off of the elevation by comparing each cell to those around it.

How steep is it, and what direction is it facing? Now we've got six environmental variables—yay! We also have spots where camas is located, so now we can go generate a habitat suitability map.

Here are some things I did to make that model, that I feel like you need to know, because they're essential ingredients in the model. One issue is resolution: the DEM is 30 meter resolution, which is a heckuva lot smaller than the PRISM data comes—they've interpolated that surface 0.00833º, which is 30 arc-seconds, which means that it is nearly a kilometer tall and about half a kilometer wide in the state of Oregon. We went ahead and downsampled everything to that lowest common denominator; introducing more uncertainty and more error in the process of meeting that criterion.

Another thing I did was to pull aside 10%, 1 in 10 (random seed, of the Oregon Flora occurrence points), to not use to train the model. I don't let the computer see them at first. I let it go through the whole modeling process and say "I made the model, I've got it," and then I throw these other points in to see how well the model did at predicting them without knowing anything about them in the first place. That's a form of Monte Carlo test, called bootstrapping (if I were to repeat this many times with different samples withheld), to see how it does.

There's one more thing that I want to point out. How generalizable is a model, as in, how generally are we drawing these lines (response curves)? The graph on the left here is a line drawn for the minimum temperature PRISM variable of the camas observations. It's a pretty wonky line—you have to do a lot of math to get there. It's a bunch of stepwise regression trees pieces of lines kind of Etch-a-Sketching their way across the variable; which is a model of your data, not the plant itself. It's a model of your sample. Compare it to the line on the right, which is one slope, and then once you hit the end of your range of variability, it doesn't matter so you just call it a flat line. This is overgeneralization. If there is a pattern there, you're probably going to miss it, because I don't know many plants that do just that. They probably have a little bit more nuance in their interaction with that variable.

These are some things to keep in mind as you think about this model, and any model you come in contact with. If a modeler is unwilling to discuss those things, ask them until they discuss them—or until they stop modeling.

Cultural centrality: humans as environmental variable

This is not very anthropological so far. I'm talking about camas because it's a culturally important plant, but I haven't done any anthropology yet. Here is the first stab at anthropology in this modeling process. I would like to introduce indigenous cultural geography, or at least a really inferential, generalized, insufficient (so far) form of it, into the model; to see if it improves; to see if it improves or changes the model's performance.

How am I going to do that? One resource at hand, which is publicly available—since I'm making this repeatable or replicable—is published culture area maps...all of which are wrong...in some way. I've never seen a human boundary that actually conforms to a crisp, vector data line across the landscape. What if we combine them, like the masses guessing the weight of an ox and the actual weight of the ox comes out of the group's average even though noone was actually right, which was a famous observation in the early 20th century. How about we combine those maps to see where they agree and disagree, then call that (very tentatively) centrality. Everyone who made a map might have thought that the same group of people was living in one same place—and put them there—but where they start to disagree, maybe those spots are more peripheral. You may thus create a center and periphery landscape. That's our new variable. Any plant can be in a very peripheral place, it can be in a very central place, or it can be somewhere in the middle.

This variable is not as fleshed out will be later, because it takes a long time to digitize culture area maps. There are the four with which I have started, and I understand that I am fudging the line between culture areas and language areas at this point. I wonder if this (animation) is going to play...yeah! I pinned each map to a standard state border, and then trace them. In this case, you'll notice

I buffered the border by 10km before I started tracing so that all the boundaries of the groups on the edge of Oregon intersect the state boundary.

The Chinook are one group which are consistent across all of these maps, so I've chosen to use them as the example. If you turn each of these bits with two polygons and give it a value of 0 where the Chinook were not and a value of 1 where the Chinook were, if you add the together then your spots with a value of 4 are, inferentially, the most central—and the spots with a value of 0 are, inferentially, the most peripheral. That's an additive process within groups, and then we're going to do a subtractive process between groups. None of the maps grouped the groups in the same way, which was really tricky, but what I ended up with was this map of centrality. It is rough, it is also wrong, but it may be closer to right.

Comparing models

Let's see if we can throw it into the model and compare it before we put the centrality variable in, and after we put it in. That's the only thing that changes—putting that centrality map into the model as an environmental variable. How are we going to do that? Two seconds to explain what these two measures of model performance are. *e* is Euler's number, it's roughly 2.74...oh man. (timekeeper for presentation has displayed the 0 minutes remaining card) Well, you can read about that on my website, since I'm at 0 minutes. (slideshow glitches and does not display table of results data) Aw, it's not going to show the data...so, the punchline is that adding centrality improves the model, and centrality ranks right in the middle of the variables in terms of variable importance; you get those by the weights of the variables as the program goes through the model. It contributes about 5.8% of the predictive power of the power of the model, which is above three and below three of the environmental variables.

Later, here's the next thing I want to do: build a couple more cultural variables, put them into the model, and quantity uncertainty...and those are my references.

Applause