Can crowd sourcing help the US census?

Noel Cressie aims to do so with a novel statistical method that employs Google Trends, Twitter and other social sites

Adam King onCampus staff

Professor Noel Cressie rides atop his noble steed, charging toward his foe. His tall, lanky body is more Don Quixote than Sir Galahad, but he exhibits no fear, only frustration that his quest for victory has eluded him thus far.

He draws his sword, focused on the blow he hopes will bring down the slithery Standard Deviation to minimize the uncertainty.

Yes, Cressie admits, this is how one statistician dreams.

"I dream about big data and how I'm going to slay them," Cressie said with a laugh. "Big data sets, they're my dragon."

Cressie's dreams are fueled by a $2.85 million National Science Foundation grant that is funding research into how to strengthen survey data. He has a two-pronged approach developed with colleagues at the University of Missouri — one an already proven method; the other a method yet to be tested. But both are aimed at benefiting the US Census Bureau's American Community Survey, which delivers annually updated estimates between the decennial census for one-, three- and five-year intervals.

The Census Bureau doesn't have the money to survey the entire populace every year to keep its information fresh. Instead, it must sample key spots and try to extrapolate those data to all 50 states. While the bureau generally provides estimates at the state level or for heavily populated metropolitan areas, it shies away from drawing conclusions at the county level because accuracy at that level is more uncertain.

But state and local governments can better act on county-level data, such as where homeowner mortgage delinquency (a mortgage payment more than 60 days late) is cropping up, where poverty rates are climbing and so on. The way to bring more certainty to the estimated values, Cressie said, is to use spatio-temporal statistics — the proven part of his team's two-pronged approach.

Noel Cressie

Spatio-temporal statistics incorporates a method called "borrowing strength," which means that counties that are nearby one another in space and time are more alike than counties that are far apart. Thus survey data from one county can be used to estimate values for an adjacent county that hasn't been surveyed.

"Certainly it has to do with the makeup of the county," Cressie said. "But in Ohio, Franklin County is more apt to look like Delaware County than Adams County or Athens County."

Out of this, two numbers emerge: One is the main data point, such as the estimated mortgage delinquency rate. The other number is the standard deviation — or what Cressie likes to call the measure of uncertainty.

"You can start to give counties estimated values and uncertainty values even though no samples were taken there," Cressie said. "Those uncertainty values might be large, but once you come up with county estimates, if it becomes important for the state to understand what is happening in certain blighted areas, it can then go in and study that area more to get further information."

When the Census Bureau publishes the aggregated three- and five-year estimates, Cressie said the bureau is confident enough to average the numbers to the county level and show the trends taking place. But Cressie and his University of Missouri colleagues, Scott Holan and Chris Wikle, are hoping to give the bureau more confidence in its annual estimates, so it can provide a data snapshot at a smaller level than merely statewide.

Cressie said he is confident that the spatial component will help take the bureau down that road, netting about 30 percent more precision for the same survey dollar. But the research group wants to be able to do even more. This is where the untested method comes in.

Cressie and his colleagues will be investigating whether using social media, and specifically crowd sourcing — in addition to traditional socio-demographic data sources — can reduce the uncertainty further.

Google Trends, Twitter and other sites with aggregate data are beginning to release their results. Using the mortgage crisis as their test case, Cressie said through specific queries they hope to be able to grab a snapshot of people who are sending out tweets or conducting mortgage-related Internet searches about what it means to have one's mortgage underwater, how to avoid foreclosure or legal issues surrounding it and so on. Though the information won't be tied to anyone's individual identity, it will identify trends by time and general location, such as a county.

If the approach works, ACS-related data could be pinpointed to show yearly trends at county levels or lower.

"The Census Bureau is aware we're looking into these extra sources of information, and they're excited," said Cressie, who spent part of his career as a Census Bureau Fellow from 1985-86 and is currently a member of the bureau's Census Scientific Advisory Committee (census.gov/cac/census_scientific_advisory_committee). "They know they don't have enough funding to do all they want to do — to make the ACS a much more useful tool so state and local governments can use it to do yearly planning in as small an area as possible.

"We might even be able to go down to census tracts, which are below the county level. It's clear initially we'll do some form of pilot study to see if these ideas are working, and we'll look at all 50 states."

The research group is hiring a postdoctoral researcher who will join the team in June, and the work of slaying dragons will continue in earnest.