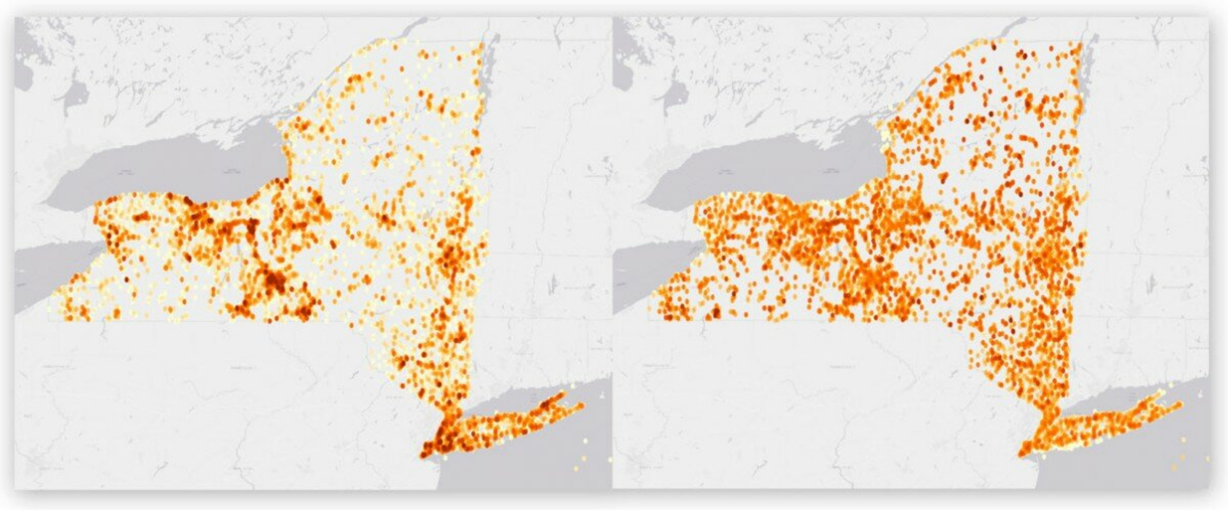
AI adjusts for gaps in citizen science data

January 28 2019, by Melanie Lefkowitz



Heat maps showing bird observations in New York state. The map on the left shows the original samples submitted to eBird and the map on the right shows the distribution after it was adjusted using a model developed by Cornell researchers to reduce location bias in citizen science projects. Credit: Cornell Lab of Ornithology

Citizen science is a boon for researchers, providing reams of data about everything from animal species to distant galaxies.

But crowdsourced information can be inconsistent. More reports come from densely populated areas and fewer from spots that are hard to access, creating challenges for researchers who need evenly distributed data.

"There is a huge bias in the data set because the data is collected by volunteers," said Di Chen, a doctoral student in computer science and first author of "Bias Reduction via End to End Shift Learning: Application to Citizen Science," which will be presented at the AAAI Conference on Artificial Intelligence, Jan. 27-Feb. 1 in Honolulu.

"Since this is highly motivated by their personal interest, the distribution of this kind of data is not what scientists want," Chen said. "All the data is actually distributed along main roads and in urban areas because most people don't want to drive 200 miles to help us explore birds in a desert."

To compensate, Chen and Carla Gomes, professor of computer science and director of the Institute for Computational Sustainability, developed a deep learning model that effectively corrects for location biases in citizen science by comparing the population densities of various locations. Gomes and Chen tested their model on data from the Cornell Lab of Ornithology's eBird, which collects more than 100 million bird sightings submitted annually by birdwatchers worldwide.

"When I communicate with conservation biologists and ecologists, a big part of communicating about these estimates is convincing them that we are aware of these biases and, to the degree possible, controlling for them," said Daniel Fink, a senior research associate at the Lab of Ornithology who is collaborating with Gomes and Chen on this work. "This gives [biologists and ecologists] a better reason to trust these results and actually use them, and base decisions on them."

Researchers have long been aware of the problems with citizen science data and have tried various methods to address them, including other types of statistical models. Projects that offer incentives to entice volunteers to travel to remote spots or search for less-popular species have shown promise, but these can be expensive and hard to conduct on a large scale.

A massive data set like eBird's is useful in machine learning, where large amounts of data are used to train computers to make predictions and solve problems. But because of the location biases, a model created with the eBird data would make inaccurate predictions.

Adjusting for bias in the eBird data is further complicated by the data's many characteristics. Each bird sighting in the system comprises 16 distinct pieces of information, making it computationally challenging.

Chen and Gomes solved the problem using a deep learning model – a kind of artificial intelligence that is good at classifying – that adjusts for population differences in different areas by comparing their ratios of density.

"Right now the data we get is essentially biased because the birds don't just stay around cities, so we need to factor that in and correct that," Gomes said. "We need to make sure the training data is going to match what you would have in the real world."

Chen and Gomes tested several models and found their deep learning algorithm to be more effective than other statistical or machine learning models at predicting where bird species might be found.

Though they worked with eBird, their findings could be used in any kind of citizen science project, Gomes said.

"There are many, many applications that rely on citizen science, and this problem is prevalent, so you really need to correct for it, whether people are classifying birds, galaxies or other situations where data biases can skew the learned model," she said.

More information: Di Chen and Carla P Gomes. Bias Reduction via End-to-End Shift Learning: Application to Citizen Science.

arxiv.org/pdf/1811.00458.pdf

Provided by Cornell University

Citation: AI adjusts for gaps in citizen science data (2019, January 28) retrieved 19 April 2024 from https://phys.org/news/2019-01-ai-adjusts-gaps-citizen-science.html