GLOBE: Analytics for Assessing Global Representativeness

Matthew D. Schmill and Tim Oates
Computer Science and Electrical Engineering
University of Maryland, Baltimore County

Lindsey M. Gordon, Nicholas R. Magliocca and Erle C. Ellis
Geography and Environmental Systems
University of Maryland, Baltimore County

***Abstract*—The goal of *meta-analysis* is to synthesize results from a collection of studies in order to identify patterns that have broader applicability. In many of the global change sciences, these synthesis studies attempt to bring together results of local case studies to make claims about global patterns. In order to substantiate claims of generality, it is crucial to establish that the collected case studies are *representative* of the regions they claim to characterize. Said differently, a meta-analyst must demonstrate that their choice of studies was not *biased* in a way that would undermine her claims. The GLOBE project aims to shorten the gap between local and global researchers by, among other things, providing analytics that help assess the representativeness of a collection of study sites and assist in correcting any bias found. In this paper we present the methods used by GLOBE to formalize the concept of representativeness, to analyze and visualize it, to address sampling bias, and present a use case in the domain of land change science.**

I. Introduction

Human learning in general, and scientific learning in particular, often proceeds by carefully generalizing from observations to potential rules or laws that govern the system being studied. These generalizations, when well-founded, facilitate accurate prediction and control. One crucial element of this process is understanding when and how far to generalize. For example, do the results of a drug trial conducted on a sample of 30 to 40 year old women say anything about the efficacy of the drug for men of the same age? What about teens, or infants? Closer to the domain of interest in this paper, can the factors that drive changes in the intensity of agricultural land use be observed and generalized across not just Central America, but across Africa as well?

At the root of questions like this is the issue of *representativeness*. Given a sample of observations and a population to which extracted rules or laws are to be applied, how similar are the sample and the population along relevant dimensions? The more similar they are, the more likely the generalizations will hold. We explore this issue in the context of case studies, each of which has its own arbitrary geography, chosen by land change scientists, where the goal is to understand the impact of human activity on the land. [1]

A discussion of the GLOBE system, a software artifact we created that facilitates integration of local case studies into global views, is followed by a precise specification three related concepts: bias, representativeness, and representedness. This is followed by two different ways of operationalizing representativeness, one based on the χ^2 statistic, and one on Monte Carlo methods. The primary challenge here is to produce a representativeness value that is formally sound, captures the intent of representativeness in the domain, and can be understood. To address the latter, visualizations of representativeness are a central part of the GLOBE system.

We present a case-study using our definitions and tools to evaluate the representativeness of a well-known meta-study of agricultural intensification in relation to human population density, and give advice on what to do if the representativeness of a collection of case studies is not sufficient. The paper concludes with a discussion and ideas for future work.

II. GLOBE

The GLOBE project is a multidisciplinary effort to transform land change science – the study of interactions between human systems, the terrestrial biosphere, atmosphere, and other Earth systems as mediated through the human use of land. The GLOBE system is a web application and collaborative environment that seeks to link local knowledge to global data to accelerate global understanding of land change processes. [2]

The central currency of GLOBE is the *case study*. Here, a case study is primary source data (such as a journal article) attached to a case geography, which describes a study site referenced in the source data. Cases have a variety of metadata associated with them, and are useful in and of themselves for descriptive purposes. The true value of a GLOBE case (or a collection of cases), though, lies in the analytical operations that can be performed by the GLOBE system's *Global Collaboration Engine* (GCE).

The GCE is the heart of the GLOBE system, providing visualization and computational tools that allow local case studies to be understood in a global context. Underlying the analytics is a data system offering (at the time of this writing) 69 global variables from a variety of categories including human factors, climate, surface features, biological, and remote sensing. These variables are processed against the GLOBE system's internal geographic representation: a *discrete global grid* (DGG) system defined using the Icosahedral Snyder Equal Area (ISEA) projection. [3] We utilize the ISEA Aperature 3 Hexagon DGG, at resolution 12, for native GLOBE data. At that resolution, there are $1,444,964$ cells that contain some land cover in the grid, which we call *Globe land units* (GLUs). Each GLU at resolution 12 has an area of approximately $95.978km^2$. An undersampled grid at resolution 10 ($160,582$ cells, area of $863.8km^2$) is also maintained for coarse visualization and fast approximated analytics.

For cases to drive the GLOBE analytical tools, their geographies must be characterized by global variable values as GLUs are. Given that a case may be described by an arbitrary geography, GLOBE approximates site global data values using an intersect-and-aggregate scheme. That is, the case geography is intersected against the GLU mesh, and the global data for the intersecting GLUs are aggregated using one of a set of selectable functions. At the time of this writing, the allowed aggregates are *mean, median*, and *centroid*. Approximations are necessary to maintain the real-time interactivity of the visualizations and analytics; in any case, aggregates are only used when case site geographies span several GLUs. We maintain that local studies should only rarely exceed this area by much, and when they do, the failure is often one of properly limiting the claims of the case study or by choosing a geometry larger than is warranted by the study. In cases where the site geometry is truly much larger than $100km^2$, adequately expressing the range of values covered requires reasoning not over data values but ranges or distributions. While there exist some analytical tools for describing the impact of choosing a large site geometry in GLOBE , It is beyond the scope of this paper to describe them here, and we will hereafter characterize case studies by their aggregate values (the default being the median [1]).

The global data layer forms a foundation on which the GLOBE system's analytics perform. At current, there are two analytical processes provided. One, the *similarity analysis*, can provide real-time visualization to highlight geographic regions that are similar to a particular case based on variable values for that case's geography, as well as the ability to search for other cases that are similar in the desired global dimensions. The other, which we focus on in this paper, is the *representativeness assessment*. This assessment also comes with real-time map-based visualizations, statistics, and search. The meaning and purpose of this analysis is described in the following section.

III. BIAS, REPRESENTATIVENESS, AND REPRESENTEDNESS

What are we really asking of a researcher when we demand that her study be representative, or, conversely, free of bias? Simply put, we are requiring that the data describing her selected case studies, taken as a sample, cannot be differentiated from a random sampling of data points taken from the global range that the claims of the meta-analysis cover. Informally, the distribution of data describing the collection of case studies should *look like* the distribution of the global range being claimed by the analyst. The goal of the work presented here is to formalize this concept.

As a start, it is useful to consider the concept of a "claimed" global range. It is rare for a meta-analyst to claim that the conclusions of their synthesis study are relevant at all points on the globe. Rather, the analyst defines a region of the earth to which their claims pertain: perhaps places where rice is cultivated, or in forested tropical areas. The GLOBE system allows the researcher to restrict their assessment based on filters defined over global variable values, and quickly see geographically what those areas look like. Careful selection of filters allows the researcher to clearly define the geographic range that their claims pertain to, and ensures that the representativeness assessment will be an accurate judge of their collection.

Once the claimed range is set, then the researcher must define one or more variables for which they claim their collection of sites to be representative. For example, one might be expected to accurately represent the *accessibility* (perhaps measured as the distance to the nearest city) of the claimed range, or, said differently, it may be a common bias to select more accessible sites in a claimed range, while under-sampling those sites that are more difficult to get to. If a large part of the claimed extent is inaccessible, then it is likely that this bias would undermine the significance of any findings in the analysis.

The representativeness of a collection of case studies, for a specified set of variables over a claimed range, then, is the degree to which the empirical probability density function (EPDF) of the collection approximates the EPDF of the GLUs in the claimed range. We will consider statistics for making this comparison in section IV. With such statistics, it is then possible to formulate a test for bias. A classical test for bias would propose the null hypothesis that the collection is drawn from the same distribution as the population of GLUs in the claimed range. From there, standard statistical hypothesis testing applies. If the null hypothesis can be rejected with a low probability of type I error, the collection is said to be biased.

Representativeness and bias are terms that characterize the collection. It is also possible to consider the degree to which areas in the claimed range are represented by the collection. This is a critical piece of the representativeness analysis that enables the user to understand geographically where their biases lie, and what to do in order to address bias. We refer to the degree to which a GLU g is represented by the collection its *representedness*. If one were to compare the EPDFs of the collection and the claimed range of GLUs, at the point represented by the global variable values of g, you would be able to characterize that difference as *over-representation* (collection has significantly higher probability density), *under-representation* (claimed range has significantly higher probability density), or *adequate representation* (collection and claimed range are not significantly different). This notion of representedness forms the basis for the GLOBE system's heat maps, as well as its analytical capability to assist in taking steps to address bias.

While the GLOBE system and its representativeness analytics are targeted at land change scientists, sampling bias is a concern in many scientific disciplines. Notably, anywhere surveys (e.g. public health surveillance [4]) or passive data gathering (e.g. social media [5]) are used to make inferences, sampling bias is a concern. Where meta-studies are concerned, there is a persistent issue of data sparsity (i.e. in the tens of data points versus thousands), and tools tailored to this case are in need, not just for land change, but all disciplines that depend on global synthesis across local studies, such as ecology, in which geographic biases in field site locations have already emerged as a serious concern. [6]

[1] The *mode* is used in the place of mean and median for categorical variables.

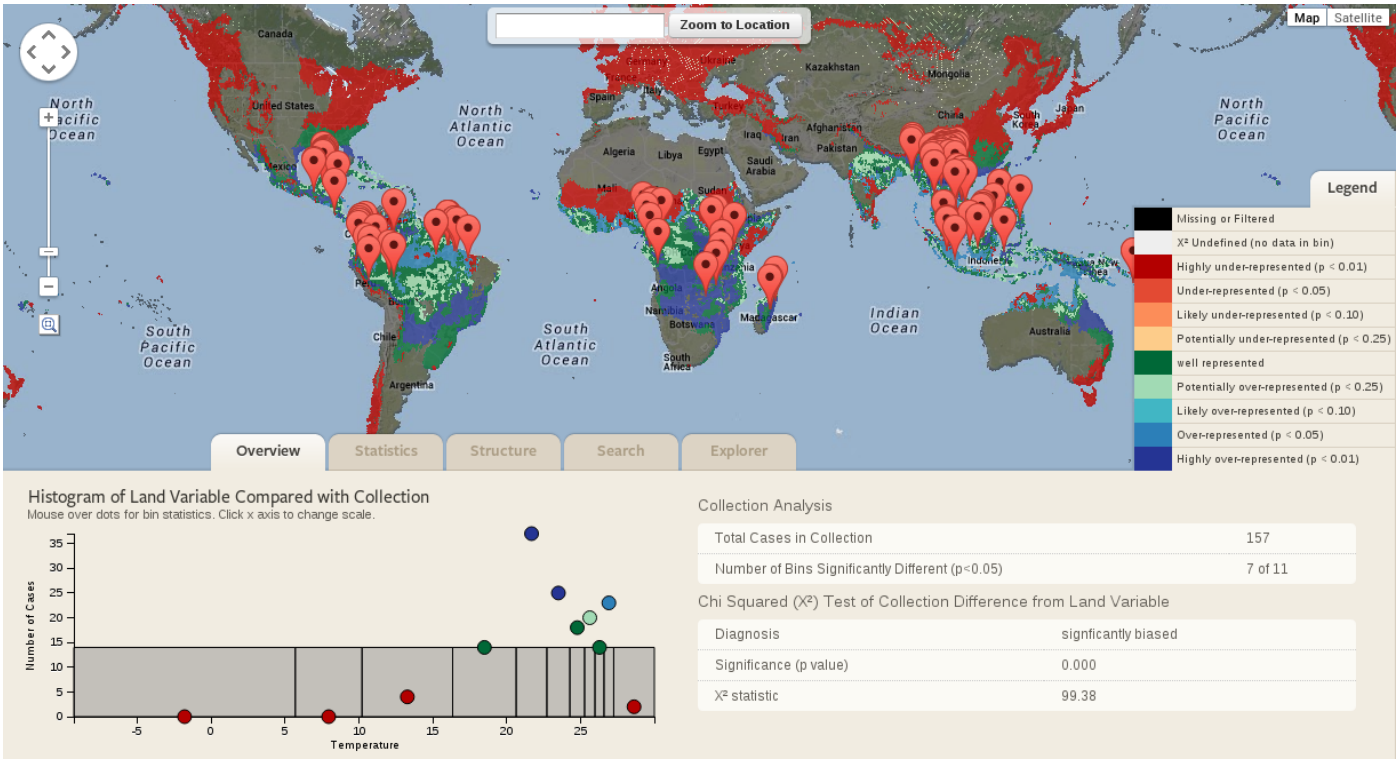

Fig. 1. A screenshot of the GLOBE representativeness assessment, featuring heat map, histograms, and χ^2 analysis.

IV. Computational Methods

With the concepts of bias, representativeness, and representedness clearly laid out, we can move on to describing the methods used by the GLOBE system. Our challenge is to deliver analytics for representativeness in real time so they can be interactive tools rather than tedious statistical processes requiring several software packages to perform. In this section, we describe three methods provided by GLOBE for assessing the various aspects of representativeness. Currently, the GLOBE system is limited to univariate analysis, and we describe the methods as univariate analysis. Multivariate analysis is planned, and we conclude this methods section with a discussion of transitioning to the multivariate case.

A. Monte Carlo Analysis

B. χ^2 Analysis

Recall the hypothesis introduced in section III that characterizes representativeness: the collection of case studies is chosen without bias from the same distribution as the claimed range. This is a suitable problem to approach with a statistical test like χ^2. Pearson's χ^2 tests for independence of two samples, using a function defined over a contingency table of observed versus expected values.

In order to utilize this statistical test, the input distributions must be discrete or made discrete. The GLOBE system provides a range of options for discretizing continuous variables including (as of this writing) equal interval, equal frequency, diagonal histogram, GLOBE custom (a custom histogram provided by the GLOBE team), and user-defined histogramming. It is a simple matter to calculate χ^2 once the variables have been made discrete, and its corresponding p value (the probability of incorrectly rejecting the null hypothesis that the collection and claimed range are drawn from the same distribution) can be used as a confidence of representativeness. The p value gives a convenient answer to the question "is my collection biased?"

The χ^2 test is also useful in computing representedness – the degree to which a particular GLU is represented in the user collection of case studies. To compute the degree of representedness for a GLU g, for global variable v, we identify the category in the discrete distribution for v into which g falls. We then run a χ^2 test for that category against all other categories in v combined. This test is computed on a simple 2×2 contingency table representing observed versus expected for both in g's category or not. We can then compute $r_v(g)$ as follows:

$$r_v(g) = \begin{cases} 0 & \text{if } 0 = f_e(g_v) = f_o(g_v) \\ (1-p) & \text{if } f_e(g_v) < f_o(g_v) \\ -(1-p) & \text{if } f_e(g_v) \geq f_o(g_v) \\ \text{undefined} & \text{if } f_e(g_v) = 0 \wedge f_0 \neq 0 \end{cases} \quad (1)$$

Where $f_e(g_v)$ is the expected frequency of the bin to which g belongs (calculated over the claimed range), $f_o(g_v)$ is the observed frequency of that bin (calculated from the user collection), and p is the p value for the χ^2 test. Note that χ^2 is undefined where the expected frequency is 0 but there are cases in the collection that occur there. [2] The range of r_g is

[2]This usually indicates that the claimed range is over-filtered as it does not even include all cases in the collection.

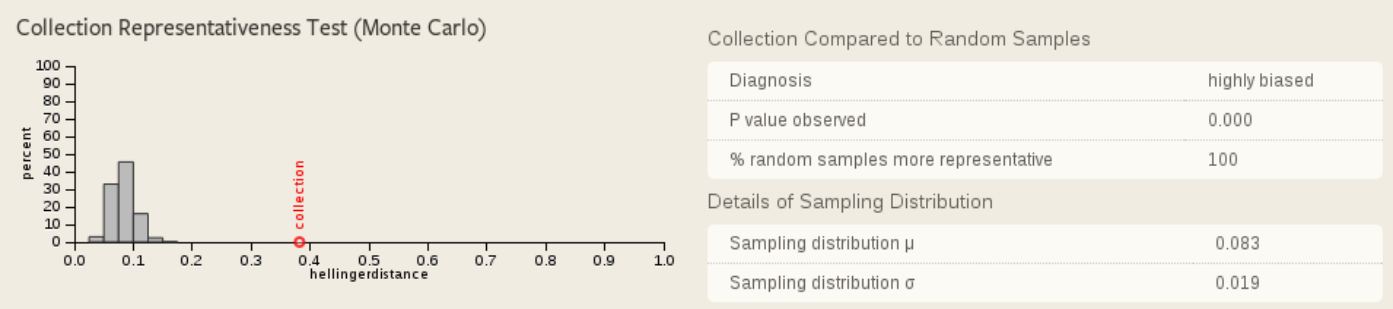

Fig. 2. A Monte Carlo bias analysis comparing a collection's Hellinger distance to random samples in the claimed range.

thus $[-1 \ldots 1]$, with 0 indicating perfect representedness, negative numbers indicating under-representedness, and positive numbers indicating over-representedness, with absolute values of 1 signifying the most extreme cases, where the probability of incorrectly concluding a biased sample is close to zero.

The resulting function r_v is suitable for rendering a map which indicates geographic regions that are under and over represented by the user collection. The default GLOBE legend for representedness goes from deep red (under-represented with $p < 0.01$) to green (well represented) to deep blue (over-represented with $p < 0.01$), with gradations occurring at typical p values $(0.05, 0.10, and 0.25)$. A screen shot of a representedness analysis is shown in figure 1. Note that this sample analysis is chosen to demonstrate the visualization; it compares a sample of tropical sites against a broad region of the globe in the dimension of average temperature. As such, there are large ranges that are under and over represented. Also note, pictured at the bottom of the screenshot, is the discretized range of temperatures, executed with an equal frequency histogram, and with collection frequencies overlaid as circles. These circles are also colored by χ^2 representedness. This is a typical starting point for representativeness assessment in the GLOBE system.

It is worth noting here that, not surprisingly, there is sensitivity in the representativeness assessment to the procedure by which continuous variables are discretized. It is possible, by gerrymandering of histograms, to influence representativeness scores, and, in extreme cases to render the results meaningless (i.e. by attempting to discretize a highly skewed variable with an equal-interval scheme). As is often the case in statistical procedures, it is up to the experimenter to conduct the assessment in earnest, and for the reader to be aware of whether the discretization has been done poorly or not. We have found that in all but extreme cases of input distribution, the equal frequency histogram produces reliable results, with little room for experimenter to adulterate the results. As such it is the default discretization strategy.

While the χ^2 analysis is standard practice in many disciplines, and produces useful results here, it is not without limitations. Chief among the concerns with this analysis is its sensitivity to small sample sizes. General recommendations with χ^2 are not to use the test with sample sizes of less than 50, or when the expected frequency for more than one category is less than 5. The statistic is also undefined where there is an expected frequency of zero. While there are alternative statistical tests (such as Fisher's exact test) that can help in these cases, another acceptable practice is the use of Monte Carlo methods to obtain a probability of bias that is less sensitive to small sample sizes.

Monte Carlo methods employ repeated random sampling to generate the distribution of a statistic whose attributes are unknown. In our case, we are concerned with statistics that compute the difference between two PDFs (a random sample and the claimed range). With such statistics, we can repeatedly take random samples, of size equal to the user collection, of GLUs in the claimed range, compute the statistic, and create an empirical distribution. We then simply compare the value of that statistic for the user's collection against the empirical distribution to derive a probability that the collection was drawn from the same distribution as the random samples (the claimed range). We describe two such classes of statistics in the following sections.

1) f-Divergence: The process of comparing probability distributions is common in a variety of computational undertakings. In information retrieval, for example, an effective technique for organizing documents by their content for effective query-answering is to form document clusters based on the similarity of their word-occurrence distributions. [7] It is common in that setting to use a metric called Kullback-Leibler Divergence, with is an information-theoretic measure of the information lost by modeling one distribution with another. [8]

Kullback-Leibler Divergence is a member of the family of f-Divergence functions that quantify the difference between two probability distributions, most of which are suitable for computing a measure of representativeness in the GLOBE system. Kullback-Leibler is not ideal in that it is asymmetric. As such, the GLOBE system offers two other functions in this family as of the time of this writing: Jensen-Shannon Divergence, and Hellinger Distance. Both are symmetric and satisfy the property that for distributions P and Q, $0 \leq f(P, Q) \leq 1$. This makes them ideal implementations for discrete representativeness in the GLOBE system, and as the default input to the Monte Carlo test for bias as a companion piece to the χ^2 tests.

A screenshot of a Monte Carlo bias analysis using Hellinger distance is shown in figure 2. The system generated 1000 random samples of the same size as the user collection (in this case, 155 GLUs in each sample), computed the Hellinger distance from the sample to the claimed range, and compared the user collection's distance against the resulting distribution. Again we see that evidence of bias in this simple example is very clear as the sample statistic is not near the empirical

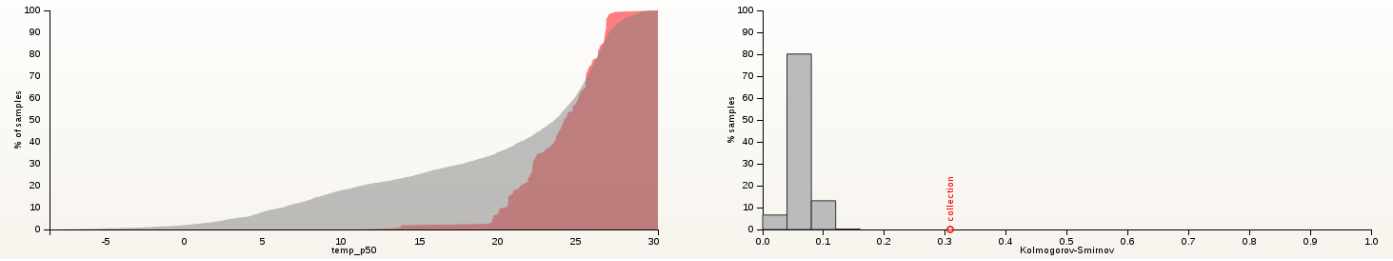

Fig. 3. A screenshot of the GLOBE system's Kolmogorov-Smirnov representativeness plugin. At left, the ECDFs for the population (gray) and collection (red). At right, the results of the Monte Carlo bias test: the K-S statistic for the user collection plotted against a sampling distribution of the K-S statistic for random samples in the claimed range.

distribution. We have found the Monte Carlo test for bias to be a useful second opinion, especially in cases of small sample size (relative to the requirements of χ^2).

2) Komolgorov-Smirnov: The sensitivity of representativeness measures to how the data range is discretized is a potential barrier to entry for users not well versed in the generation of meaningful and useful histograms. For this reason, we have investigated methods that do not require the step of making the population PDF discrete. A particularly interesting example is the Kolmogorov-Smirnov test [9], a goodness-of-fit test that compares the empirical cumulative distribution functions (ECDFs) of two samples. The maximum difference is taken to be the test value, and critical values have been published. [10] The Monte Carlo test for collection representativeness can also be computed with Kolmogorov-Smirnov as the test statistic.

A screenshot of the Kolmogorov-Smirnov plugin to GLOBE is shown in figure 3. Pictured are the overlaid ECDFs for the population and collection (at left - the population is gray and the sample is red), and the result of a Monte Carlo test using the two sample K-S statistic as the test statistic. Again, in this simple example generated to demonstrate bias, it is clear that bias is present. At left, the ECDF overlay evidences bias with a significant difference in smaller values of the global variable (i.e., the sample contains far fewer cold places than the claimed range). At right, the Monte Carlo analysis demonstrates bias as the collection statistic is fair from the range of random samples found in the empirical distribution.

The Kolmogorov-Smirnov test is a useful test for judging the representativeness of a sample without the requirement that the input distributions be discretized. However, this method comes with some caveats. The first is that a direct application of the statistic to computing representedness has not been developed, and thus generating a heat map is an open concern. The second is that transitioning the K-S test to multiple dimensions can be challenging, both conceptually and computationally, especially beyond two dimensions. [11] Though we only present univariate analysis in this paper, the transition to multivariate analyses is forthcoming. Our thoughts on that subject follow.

C. Multivariate Analysis

While we have concentrated on univariate analysis, there is little practical difficulty (with the exception of Kolmogorov-Smirnov) in extending the methods described here to multivariate cases. There are, however, significant challenges in making the multivariate case intuitive, presentable, and foolproof for the user. Central to these challenges is "scaling up" the core visualization of the univariate analysis: the histogram. The overlaid histogram provides the most natural and understandable explanation of the representativeness concept.

A common approach to reducing the complexity of mutlidimensional analysis is to reduce the dimensionality of the inputs using an technique like *principal component analysis* (PCA). [12] While this would allow simple, univariate analyses and visualizations to be used, our findings were that the PCA reduction broke the relationship between the variable input and the representativeness output, and simple cases could be generated where a representative collection could be judged unrepresentative (and vice-versa). Thus the analysis results were no longer meaningful, and the obvious choice for dimensionality reduction cannot be used.

The alternative is to remain in unreduced variable space. In spite of our inability to visualize them, multidimensional histograms can be used and the methods described in section IV-B are applicable. One must be cautioned, however, that as the number of bins increases, the expected values in those bins tend to decrease, and so increases the need for many source studies to properly cover the many conditions (histogram bins) defined by the joint distribution. We are investigating metrics for relating the number of bins in the discretized joint distribution to the number of study sites required to achieve an acceptable representativeness score. Multivariate analysis continues to be an ongoing topic of research.

V. REMEDIATION OF BIASED COLLECTIONS

Once bias has been detected in a collection of study sites, the researcher has two options for addressing the issue. The first (preferred) mechanism for addressing sample bias is to add studies that address under-represented regions of the specified variable space and remove studies from over-represented regions. We refer to this process as *sample fortification*. The alternative is to assign additional weight to studies in under-represented areas and reduce the weight of studies in over-represented areas in the metastudy. This process is called *reweighting*.

There are many approaches to reweighting (e.g. [13]); it can be accomplished either in closed-form (discrete analyses) or by gradient descent method to arrive at a set of weights that maximizes representativeness. An extremely simple weighting mechanism that is available in GLOBE is to weight each case

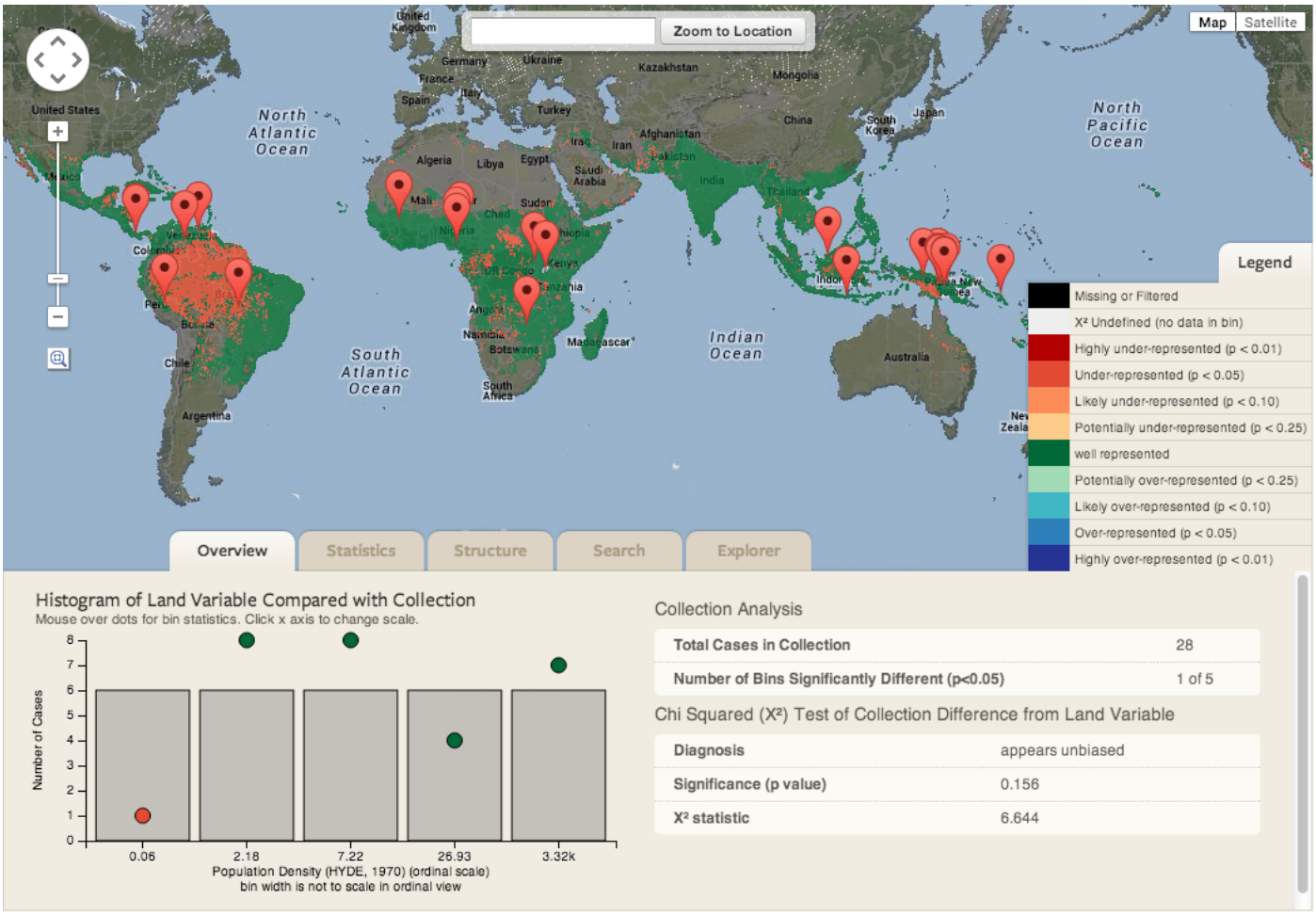

Fig. 4. Turner, et al. 1977 use case: representedness map, overlaid histogram, and χ^2 test results for the collection as found in the original meta-study.

c as $\frac{p_p(b)}{p_c(b)}$, where $p_p(b)$ is the probability mass for the claimed range in bin b where c falls into b, and $p_c(b)$ is the probability mass for the collection in bin b where c falls into b. This weight can then be normalized so the sum of weights equals the number of cases. Reweighting is is useful for producing intermediary or preliminary results (as fortification is more time-consuming), and in situations where there is a dearth of available studies to fortify an under-represented range.

Sample fortification is recommended for addressing bias when it is possible, as it works by incorporating additional data rather than boosting the importance of existing data points relative to others. Sample fortification is facilitated in the GLOBE system by *representedness search*. Study gaps (under-represented areas) are addressed one at a time, prioritized by the level of under-representedness, with a full-featured search of the GLOBE case database. Search queries may include full-text, metadata, and geographic constraints, and results can be ordered by the same features. The workflow for fortification in GLOBE is:

1) **Assess** representativeness
2) **Search** for cases that address bias
3) **Add** to the collection
4) **Repeat**

Most researchers will employ a strategy that includes first sample fortification, then reweighting, to arrive at a collection of studies that is demonstrably representative.

VI. A Practical Example

To illustrate the GLOBE systems analytical capabilities, we provide a use case of a representativeness analysis, including assessment and remediation of bias in a classic study of land-use intensification. We created a collection of 28 cases in GLOBE based on a 1977 Turner, Hanham, and Portararo meta-analysis that examines the relationship between agricultural intensification and population density in tropical subsistence communities. [14] Of the 29 case studies included in the original publication, one case was excluded due to unavailable source data, and the remaining 28 cases were analyzed and georeferenced into the GLOBE system as a collection of point and non-point geometries.

Since the objective of the meta-analysis is to examine the trajectory of agricultural intensification as it relates to population density in areas suitable for tropical subsistence agriculture, we analyzed how well Turner et al.s collection captures the range of variation in a global set of 1970 population density data. [15] In order to adequately illustrate the global land area suitable for tropical subsistence agri-

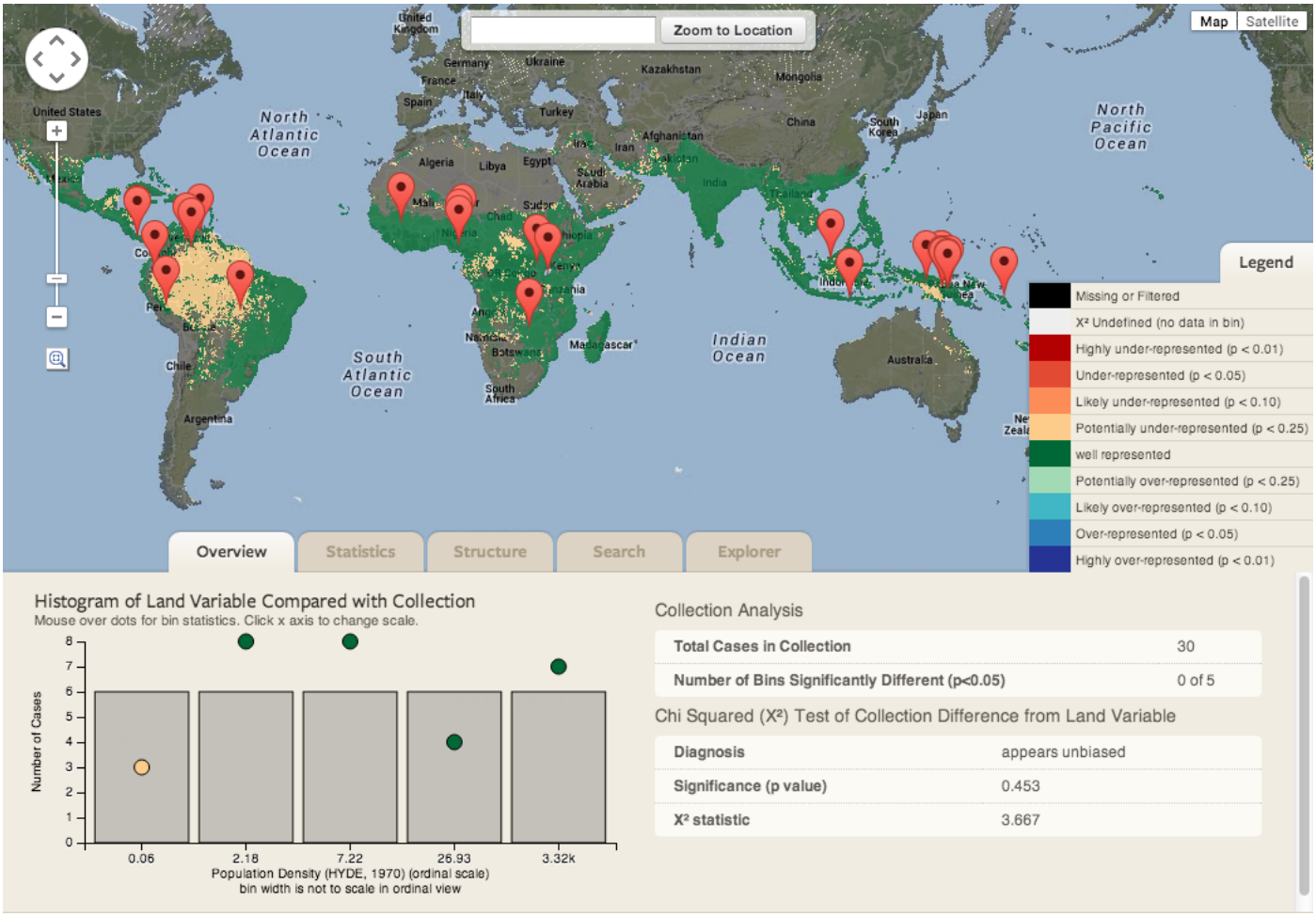

Fig. 5. Turner, et al. 1977 use case: representedness map, overlaid histogram, and χ^2 test results after sample fortification to address potential biases in the original collection.

culture, we filtered the global extent of the analysis using mean annual precipitation above 60 mm [16], mean annual temperature above 18° C [17], tropical ecoregions[18], and anthropogenic biomes to exclude wildlands, urban areas, and pastoral regions [19]. Population density was discretized using the Equal Frequency binning strategy with 5 bins, producing the results seen in figure 4. The χ^2 analysis diagnoses produces a weak rejection of the hypothesis that there is bias present ($p = 0.156$), which is confirmed by the Monte Carlo analysis (not pictured, $p = 0.062$). Visual inspection of the heat map, which features a geographic range colored orange (likely under-represented), and the χ^2 histogram, which contains the corresponding bin where the collection count is clearly than the that of the claimed range, suggests that regions with low population density may have been undersampled.

In order to address this potential bias, we used the search function of GLOBE to add more cases in undersampled areas. Using subsistence and shifting cultivation keywords, a search through the GLOBE system produced two relevant cases in the desired population density range. These were added to fortify the original collection. The representativeness analysis was conducted on the revised collection (N=30), and results can be seen in figure 5. The previously undersampled areas have been improved with the addition of more case studies, and the χ^2 test reports a p-value more consistent with an unbiased sample ($p = 0.453$). The Monte Carlo assessment shown in figure 6 confirms that the revised collection is statistically indistinguishable from an unbiased sample $p = 0.426$. Any remaining discrepancies (though they are not deemed significant) between the collection and global distributions will be accounted for by reweighting each case to maximize representativeness. Case weights may be exported by the GLOBE system and used in statistical analysis to further investigate the relationship between population density and agricultural intensity in Turner et al.s collection of studies.

VII. CONCLUSIONS

This paper is an attempt to formalize the concepts of representativeness and representedness for the purposes of evaluating and addressing sampling bias in meta-analysis. Our project, called GLOBE , implements a variety of methods for making assessments about the representativeness of a collection of case studies that is tailored to the Land Change Science community. Methods of testing for bias include classical statistical hypothesis testing (χ^2) and Monte Carlo tests using an information-theoretic approach (f-divergence) to representativeness as well as a non-parametric test (Kolmogorov-Smirnov). Those tests, paired with a measure of representedness, combine to provide

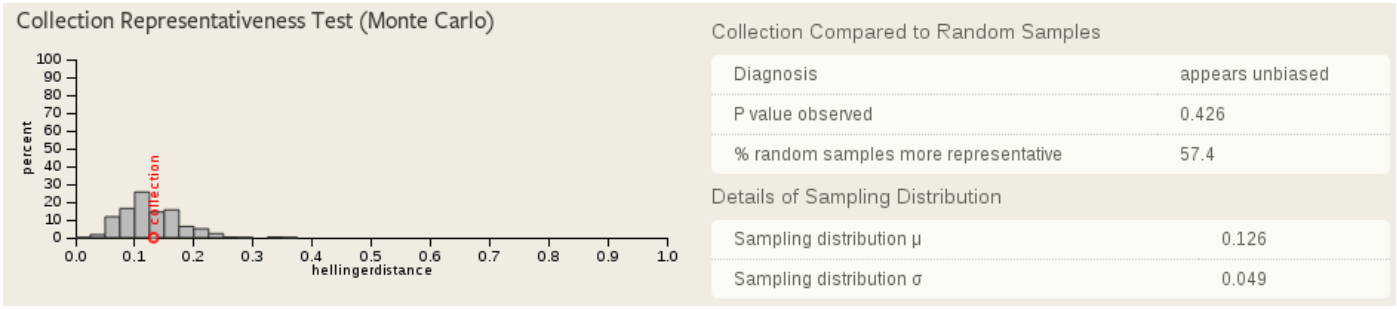

Fig. 6. A Monte Carlo representativeness bias assessment after sample fortification of cases using Hellinger distance

a suite of real-time visualization, exploratory, and statistical tools aimed at allowing researchers to do better science. We presented a actual use case of the tool in revisiting a highly influential land change paper, and demonstrated how the tools could be used to assess potential bias in the studies it comprised, and then make changes to arrive at a more representative set of studies.

We have identified caveats with the various metrics used in this analysis, and also addressed the complexity of transitioning to multivariate representativeness. Our plan is to address both in future work. While land change science is currently our domain of interest, it just one discipline that can benefit from the concepts presented here, and we are optimistic about the prospects of enabling GLOBE for a variety of disciplines.

ACKNOWLEDGMENTS

The authors would like to thank Patrick O'Connell for his contributions to the Globe GCE.

This material is based upon work supported by the US National Science Foundation under grant NSF # 1125210, and cosponsored by the Global Land Project (GLP; www.globallandproject.org). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] N. R. Magliocca, T. K. Rudel, P. H. Verburg, W. J. McConnell, O. Mertz, K. Gerstner, A. Heinimann, and E. C. Ellis, "Synthesis in land change science: methodological patterns, challenges, and guidelines," *Regional Environmental Change*, pp. 1–16, 2014. [Online]. Available: http://dx.doi.org/10.1007/s10113-014-0626-8

[2] A. L. Young, W. G. Lutters, N. R. Magliocca, and E. C. Ellis, "Designing a system for land change science meta-study," in *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2013, pp. 1473–1478.

[3] K. Sahr, D. White, and A. J. Kimerling, "Geodesic discrete global grid systems," *Cartography and Geographic Information Science*, vol. 30, no. 2, pp. 121–134, 2003. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1559/152304003100011090

[4] B. C. Choi and A. W. Pak, "Understanding and minimizing epidemiologic bias in public health research," *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique*, pp. 284–286, 2005.

[5] F. Morstatter, J. Pfeffer, and H. Liu, "When is it biased? assessing the representativeness of twitter's streaming api," *CoRR*, vol. abs/1401.7909, 2014.

[6] L. J. Martin, B. Blossey, and E. Ellis, "Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations," *Frontiers in Ecology and the Environment*, vol. 10, no. 4, pp. 195–201, May 2012. [Online]. Available: http://dx.doi.org/10.1890/110154

[7] J. Xu and W. B. Croft, "Cluster-based language models for distributed retrieval," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 254–261. [Online]. Available: http://doi.acm.org/10.1145/312624.312687

[8] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 03 1951. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177729694

[9] F. J. Massey, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951. [Online]. Available: http://www.jstor.org/stable/2280095

[10] P. E.S. and H. Hartley, "Biometrika tables for statisticians, volume 2," *Technometrics*, vol. 2, pp. 117–123, 1972.

[11] R. H. C. Lopes, P. R. Hobson, and I. D. Reid, "Computationally efficient algorithms for the two-dimensional kolmogorovsmirnov test," *Journal of Physics: Conference Series*, vol. 119, no. 4, p. 042019, 2008. [Online]. Available: http://stacks.iop.org/1742-6596/119/i=4/a=042019

[12] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psych.*, vol. 24, 1933.

[13] J. Sanchez-Meca and F. Marn-Martnez, "Weighting by inverse variance or by sample size in meta-analysis: A simulation study," *Educational and Psychological Measurement*, vol. 58, no. 2, pp. 211–220, 1998. [Online]. Available: http://epm.sagepub.com/content/58/2/211.abstract

[14] B. Turner, R. Q. Hanham, and A. V. Portararo, "Population pressure and agricultural intensity," *Annals of the Association of American Geographers*, vol. 67.3, pp. 384–396, 1977.

[15] K. K. Goldewijk, A. Beusen, and P. Janssen, "Long-term dynamic modeling of global population and built-up area in a spatially explicit way: Hyde 3.1," *The Holocene*, vol. 20, no. 4, pp. 565–573, 2010. [Online]. Available: http://hol.sagepub.com/content/20/4/565.abstract

[16] C. Willmott and K. Matsuura, "Terrestrial air temperature and precipitation: Monthly and annual climatologies (version 3.02)," Center for Climatic Research, Department of Geography, University of Delaware, 2001.

[17] E. Girvetz, C. Zganjar, G. Raber, E. Maurer, and e. a. Kareiva, P, "Applied climate-change analysis: The climate wizard tool," *PLoS ONE*, vol. 4, no. 12, 2009.

[18] D. M. Olson, E. Dinerstein, E. D. Wikramanayake, N. D. Burgess, G. V. N. Powell, E. C. Underwood, J. A. D'amico, I. Itoua, H. E. Strand, J. C. Morrison, C. J. Loucks, T. F. Allnutt, T. H. Ricketts, Y. Kura, J. F. Lamoreux, W. W. Wettengel, P. Hedao, and K. R. Kassem, "Terrestrial ecoregions of the world: A new map of life on earth," *BioScience*, vol. 51, no. 11, pp. 933–938, nov 2001. [Online]. Available: http://dx.doi.org/10.1641/0006-3568(2001)051\%5B0933:teotwa\%5D2.0.co;2

[19] E. C. Ellis and N. Ramankutty, "Putting people in the map: anthropogenic biomes of the world," *Front Ecol Environ*, vol. 6, no. 8, pp. 439–447, 2008. [Online]. Available: http://www.ecotope.org/people/ellis/papers/ellis_2008.pdf