International Cartographic Association
Association Cartographique Internationale

Does Your Mapped Variable Add Up? Statistical Guardrails for Web-Based Cartogram Generation

Adi Singhania [a], Atima Tharatipyakul [a], Nihal Z. Miaji [b], Michael T. Gastner [a,*]

[a] *Information and Communications Technology Cluster, Singapore Institute of Technology, Singapore, aditya.singhania@singaporetech.edu.sg, atima.tharatipyakul@gmail.com, michael.gastner@singaporetech.edu.sg*
[b] *Mathematical, Computational and Statistical Sciences, Yale-NUS College, Singapore, nihalzp9@gmail.com*

* Corresponding author

Keywords: web cartography, thematic maps, extensive data, machine learning

Abstract:

Area cartograms (i.e., maps that visualize statistical quantities by resizing geographic regions) are generally interpretable only when area represents an additive total rather than a derived quantity such as a density, rate, proportion, index, or percentage change (Golebiowska et al., 2021). However, analyzing server logs of go-cart.io, a web application for generating cartograms (Shi et al., 2019), we found that 23% of cartogram requests involve variables that are not spatially extensive (i.e., not additive over disjoint regions, as defined by Top et al., 2022) and, therefore, do not represent totals. To address this problem, we integrated a statistical screening tool into go-cart.io that warns users when the uploaded variable is unlikely to be extensive and explains why the input may be unsuitable.

The screening model translates extensivity into a scaling question: holding population density fixed, do larger regions tend to have proportionally larger expected values? For extensive variables, such as gross domestic product (GDP) or CO_2 emissions, doubling a region's area should approximately double the expected value. More generally, we fit an area-scaling relationship to the data: if area doubles, the expected value changes by a factor of 2^{γ}. Extensivity implies $\gamma = 1$.

To estimate γ, we treat the observed value for each region in a cartogram (e.g., an administrative division) as the outcome of a random process with mean μ that depends only on area A and on intensive variables (i.e., properties independent of system size). For this pilot study, we assume that population density ρ is the only relevant intensive variable. Although restrictive, this assumption provides a plausible starting point because many socially and economically relevant intensive variables are strongly associated with ρ.

To test extensivity, we fit the generalized additive model $\log \mu = \gamma \log A + \beta_0 + s(\log \rho)$, where β_0 is a constant and s is a regression spline; log-transforms are used because many geospatial variables are right-skewed. We assess evidence against extensivity by calculating a one-sided confidence interval $(-\infty, \gamma_{\max})$ for the area-scaling coefficient γ at significance level $\alpha = 0.05$. A value of $\gamma_{\max}$ below 1—the value required by extensivity—indicates that the variable is unlikely to be suitable for area encoding in a cartogram.

To evaluate the screening rule, we assembled benchmark data on area, population density, and 34 diverse economic and environmental variables measured across first-level administrative divisions (e.g., states or provinces). The variables were evenly split between 17 extensive quantities (including GDP and CO_2 emissions) and 17 non-extensive quantities (including fraction of forest cover and mean travel time to the nearest city). For each variable, we included only countries with at least 10 non-missing division-level observations, yielding 1,717 extensive and 1,920 non-extensive country-variable combinations. Using R's mgcv package, we fitted generalized additive models with thin-plate spline bases of dimensions 4, 6, and 8. A warning is issued only if all fitted models yield $\gamma_{\max}$ below the warning threshold. Varying the threshold produced an ROC curve with an AUC of 0.967. We selected $\gamma_{\max} < 0.8$ as the warning criterion, yielding a false-positive rate of 3.5% and a false-negative rate of 6.6%. Data and R scripts are available at figshare (Singhania et al., 2026).

We integrated this statistical model into go-cart.io (version 4.6.0). After users upload data and request a cartogram, the system evaluates the input against the model. An advisory warning—shown in Figure 1—is displayed only when a potential extensivity violation is detected, thereby reducing unnecessary warnings and the risk of user fatigue. To facilitate verification, go-cart.io then asks whether merging two specific regions—chosen to be large and to differ substantially in mapped values—should result in the sum of these values or an intermediate value. This dialog is strictly advisory; users may proceed if they deem their data appropriate for the chosen visualization.

https://doi.org/10.5194/ica-abs-12-126-2026

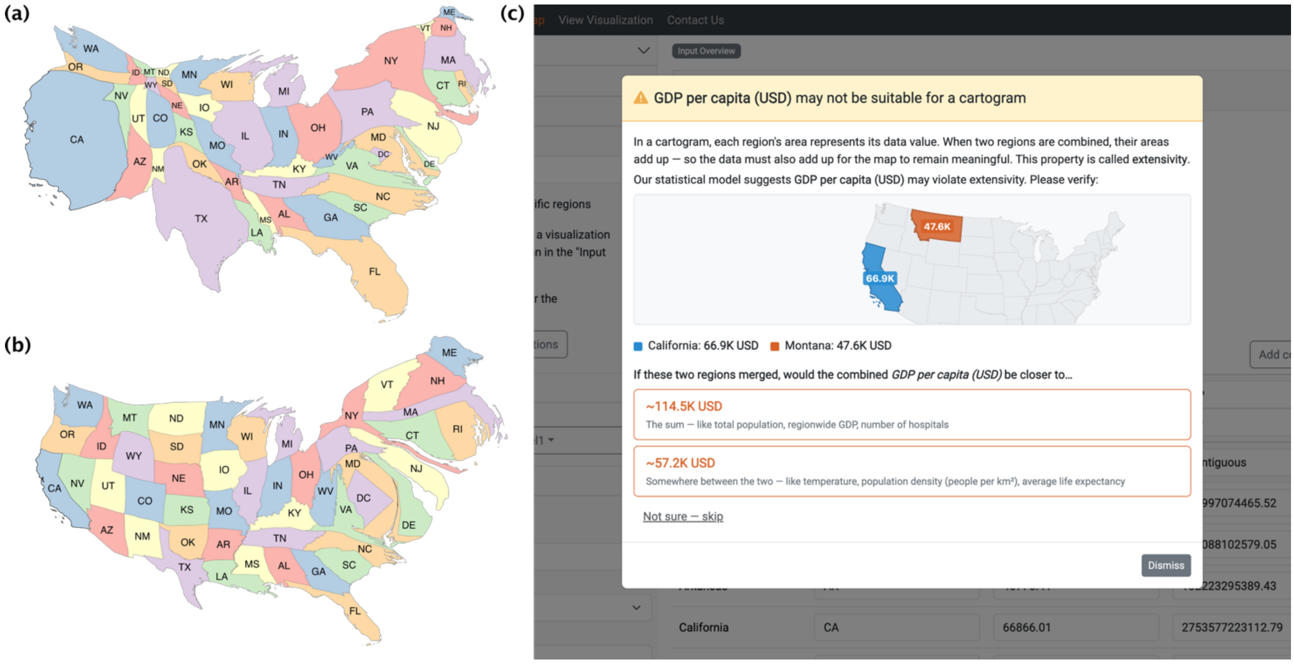

Figure 1. Cartograms of states in the contiguous USA and the go-cart.io non-extensivity warning. (a) Cartogram with state areas in proportion to gross domestic product (GDP). Because GDP is an extensive variable, this cartogram provides a meaningful visualization. (b) GDP per capita (an intensive variable) mapped as a cartogram. This cartogram is misleading because state areas do not represent fractions of a total. (c) When the statistical model detects a potential extensivity violation, a warning dialog explains why cartograms require extensive data and asks the user to verify whether the variable is additive.

The statistical guardrail is intended to reduce the proportion of misleading cartograms produced by non-expert users—an important consideration as web cartography tools reach broader audiences. However, several limitations remain. First, the current model assumes that population density is the only relevant intensive variable governing the scaling relationship; incorporating additional variables (e.g., urbanization rate) and historical data might improve the model's discriminatory power. Second, the system warns against non-extensive variables but currently does not recommend an alternative map type (e.g., a choropleth map), leaving that decision to the user. Future work should address these limitations, assess the model's generalizability to user-uploaded data, and evaluate whether the interactive verification step measurably reduces non-extensive cartogram submissions.

Acknowledgements

This project was supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (EP2) programme (Award No. MOE-T2EP20221-0007). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Ministry. We acknowledge the use of Claude 4.6 for code development and ChatGPT 5.4 for final language polishing. The research design and interpretation, however, are solely the work of the authors.

References

Golebiowska, I., Korycka-Skorupa, J. and Slomska-Przech, K., 2021. Common thematic map types. In: Wilson, J.P. (Ed.), The Geographic Information Science & Technology Body of Knowledge, 2nd Quarter 2021 Edition. DOI: 10.22224/gistbok/2021.2.7

Shi, T., Duncan, I.K. and Gastner, M.T., 2019. go-cart.io: A web application for generating contiguous cartograms. Abstracts of the International Cartographic Association, 1, 333. DOI: 10.5194/ica-abs-1-333-2019

Singhania, A., Tharatipyakul, A., Miaji, N.Z. and Gastner, M.T., 2026. Does Your Mapped Variable Add Up? Statistical Guardrails for Web-Based Cartogram Generation. figshare. Dataset. DOI: 10.6084/m9.figshare.31915695.v1

Top, E., Scheider, S., Xu, H., Nyamsuren, E. and Steenbergen, N., 2022. The semantics of extensive quantities within geographic information. Applied Ontology, 17(3), pp. 337–364. DOI: 10.3233/AO-220268

https://doi.org/10.5194/ica-abs-12-126-2026