How and why do some authors claim their data and models are good?

Robert Gilmore Pontius Jr

Professor, Clark University

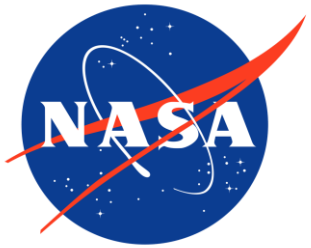
PhD, State University of New York / College of Environmental Science and Forestry

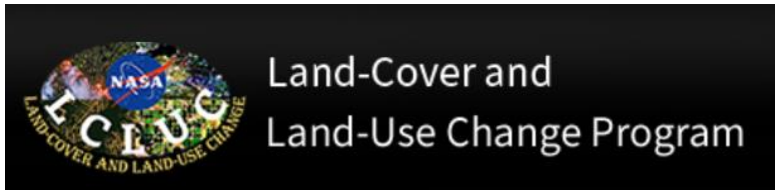
Master of Applied Statistics, The Ohio State University

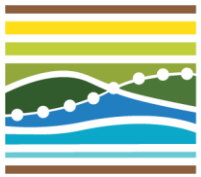
www.clarku.edu/~rpontius

rpontius@clarku.edu

12 June 2024







NATIONAL SCIENCE FOUNDATION
LTER NETWORK
LONG TERM ECOLOGICAL RESEARCH

MAPBIOMAS

Conclusions and Recommendations

1. Our profession has enormous problems, many of which are social problems related to our motivations, which respond to problematic incentive systems.
2. You can advance your career, keep your dignity, and enjoy your work, but you must be willing to report what you see.
3. Listen to students such as Orsi Varga, Claudia Viana, Aiyin Zhang, and Thomas Bilintoh.
4. Read the book Metrics That Make a Difference starting with the chapter entitled “Commandments to Avoid Deadly Sins”.

How and why do some authors claim their data and models are good?

1. Authors routinely tell audiences that their data or models are good and that the authors are working to make them better.
2. Pontius asks the authors to define what they mean by good and better. Furthermore, if the data and models are good then why do the authors want them to be better. These are sincere questions, which Pontius has been asking for decades.
3. Many people have told Pontius that they fear when he is at their presentations because of the questions he usually asks. Other people are inspired by the questions.
4. Pontius is not saying that models are bad or that their methods are flawed, although they frequently are.
5. Pontius is sincerely interested in the authors' motivations and how those motivations influence the next steps in their research agendas.

Here are responses when Pontius asks authors to define what they mean by *good*.

Typical responses indicate at least one of the following:

1. The authors use a flawed metric.
2. The authors use a flawed threshold of the metric.
3. The authors use popular metrics because the authors assume popularity implies validity.
4. The authors equate good with more detailed and mathematically complicated.
5. The authors are following the guidance of their professors.
6. The authors cannot express any coherent reason for making their claim.
7. The authors main goal is to publish a journal article to advance their careers.

Rare responses include:

1. The goal of the model is to facilitate a conversation among stakeholders.
2. The behavior of the model matches the modeler's intuition of the way the world works.

Table 3.3. *Summary of some of the motivations that people and organizations have in undertaking quantitative policy-focused research and analysis*

Substance-focused motivations

1. To obtain an "answer" to a specifically formulated policy question that will be directly implemented without further consideration or interpretation.
2. To develop insight and understanding that will be useful to one or several policymakers who are faced with making decisions on a specific well defined set of policy issues or options.
3. To illuminate and provide insight on a general area of policy concern for a variety of interested parties.

Position-focused motivations

4. To provide substantiation and arguments to support one's views in an adversarial procedure.
5. To generate an answer in situations in which one is expected to justify the action taken on the basis of the scientific and technical specifics of the problem.

Process-focused motivations

6. To persuade others that one has got things under control, knows what he or she is doing, and should be trusted.
7. Because the law says one must.
8. Because other people expect it.
9. Because it is not clear what else to do and the political reality of the situation requires that one do something.

Analyst-focused motivations

10. To derive enjoyment and professional recognition and rewards.
11. To use a specific problem context as a vehicle to demonstrate, test, refine, or develop new analytical techniques and tools.
12. Because it is the analyst's job.
13. Because it is the only thing the analyst knows how to do.

Researchers have several of these motivations.
People tell Pontius that their motivation is substance-focused (3), but Pontius sees that their actions are focused mainly on Position (5), Process (6), and Analyst (10).

Morgan and Henrion (1990) UNCERTAINTY: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press.

Pop Quiz

Does Comparison 1 or Comparison 2 agree more with the Reference?

Reference

1	0	0	0	0
0	0	0	0	0

Comparison 1

1	0	1	1	1
1	1	1	1	1

Comparison 2

0	1	0	0	0
0	0	0	0	0

Options for answers are

Comparison 1

Comparison 2

Other

Pop Quiz

Kappa says Comparison 1 agrees more with the Reference?

Reference				
1	0	0	0	0
0	0	0	0	0

Comparison 1				
1	0	1	1	1
1	1	1	1	1

Comparison 2				
0	1	0	0	0
0	0	0	0	0

Kappa

(0.20-0.18)/(1-0.18) ≈ 0.02

(0.80-0.82)/(1-0.82) ≈ -0.11

Pontius and Millones (2011) published the *Death To Kappa* paper, which had two messages:

Don’t use Kappa.

Use quantity and allocation disagreement.

The *Death to Kappa* paper has more than 1900 citations.

Pontius and Millones (2011) Death to Kappa. International Journal of Remote Sensing. https://www.tandfonline.com/doi/abs/10.1080/01431161.2011.552923

Abstract of the *Death to Kappa* paper

This article concludes that these Kappa indices are useless, misleading and/or flawed for the practical applications in remote sensing that we have seen. After more than a decade of working with these indices, we recommend that the profession abandon the use of Kappa indices for purposes of accuracy assessment and map comparison, and instead summarize the cross-tabulation matrix with two much simpler summary parameters: quantity disagreement and allocation disagreement.

Here is how some authors cite the *Death To Kappa* paper by Pontius and Millones (2011)

"kappa coefficient … has proved to be an excellent statistical parameter for measuring consistency (Pontius and Millones 2011)."

cited in Gao et al. (2021) https://doi.org/10.1016/j.ijdrr.2020.101928

Half of the manuscripts that cite the *Death to Kappa* paper still use Kappa.

Kappa is popular in the assessment of predictive models of land change.

Many authors still use kappa for model validation because the software automatically computes kappa.

Zhang et al. (2021) obtained a kappa of 0.88 for a simulation model and claimed that "**Kappa coefficients greater than 0.8 indicate that the simulation accuracy has reached an ideal state [79]**" where citation [79] from the prestigious *International Journal of Geographical Information Science* does not even contain the word kappa.

Two-map validation of a prediction shows proportion correct is 0.70 and kappa is 0.67 for ten classes

Time 2 Prediction

1 X	2	3 X	4 X	5
6	7	8	9	10

Time 2 Reference

3	2	4	1	5
6	7	8	9	10

Two-Map Comparison of Prediction versus Time 2 Reference

Kappa = [0.7 - (10 X 0.1 X 0.1)] / [1 - 0.1] = 0.6/0.9 ≈ 0.67

The two-map comparison is at a time point, not during a time interval.
Assessment at a time point cannot evaluate change during a time interval.

Three-map validation of a prediction of change shows more error than hits. A null model of zero change also has kappa = 0.67.

Time 1 Reference

1	1	1	1	5
6	7	8	9	10

Time 2 Prediction

1	2	3	4	5
6	7	8	9	10

Time 2 Reference

3	2	4	1	5
6	7	8	9	10

Three-Map Comparison

M	H	W	F	C
C	C	C	C	C

Two-Map Comparison of Prediction versus Time 2 Reference

Kappa = [0.7 - (10 X 0.1 X 0.1)] / [1 - 0.1] = 0.6/0.9 ≈ 0.67

Two-Map Comparison of Time 1 versus Time 2 Reference

Kappa = [0.7 - (0.04 + 6 X 0.1 X 0.1)] / [1 - 0.1] = 0.6/0.9 ≈ .67

The three-map comparison shows helpful assessment in terms of Misses (M), Hits (H), Wrong Hits (W), False Alarms (F), and Correct Rejections (C).

We invited land-change modelers to submit:

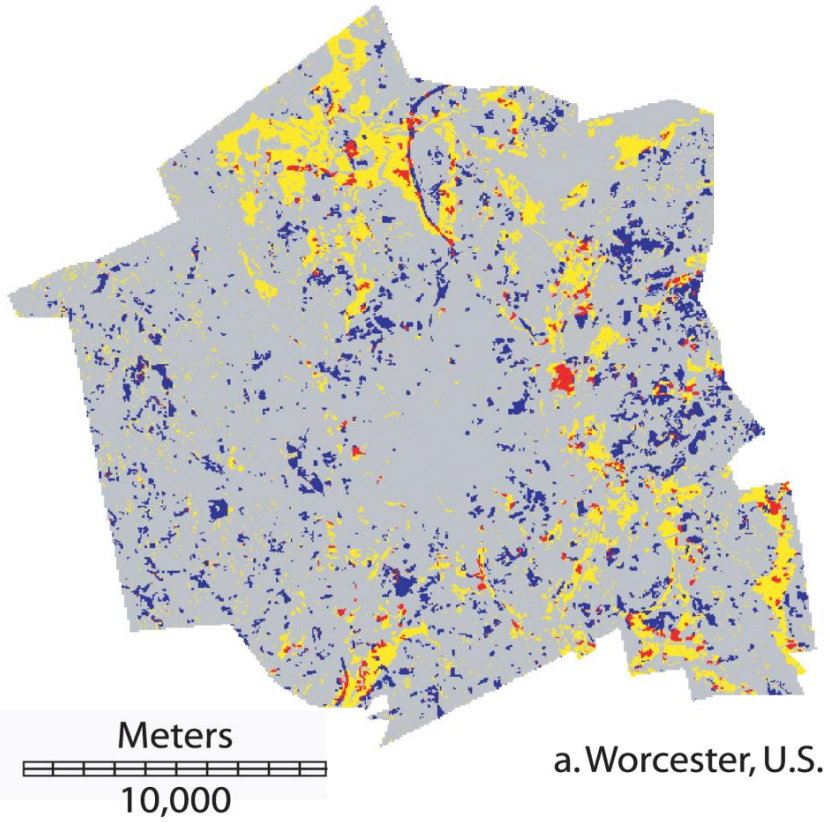
1. Reference Map of Time 1,
2. Reference Map of Time 2,
3. Prediction Map of Time 2,
4. Criterion to evaluate the maps.

Pontius Jr et al. 2018. Lessons and Challenges in Land Change Modeling Derived from Synthesis of Cross-Case Comparisons. Chapter 8 in Martin Behnisch and Gotthard Meine (eds.) Trends in Spatial Analysis and Modelling. Geotechnologies and the Environment 19: 143-164. Springer International Publishing: Cham, Germany.

Twelve sites of thirteen applications

Pontius Jr et al. 2018. Lessons and Challenges in Land Change Modeling Derived from Synthesis of Cross-Case Comparisons. Chapter 8 in Martin Behnisch and Gotthard Meine (eds.) Trends in Spatial Analysis and Modelling. Geotechnologies and the Environment 19: 143-164. Springer International Publishing: Cham, Germany.

Geomod by Pontius



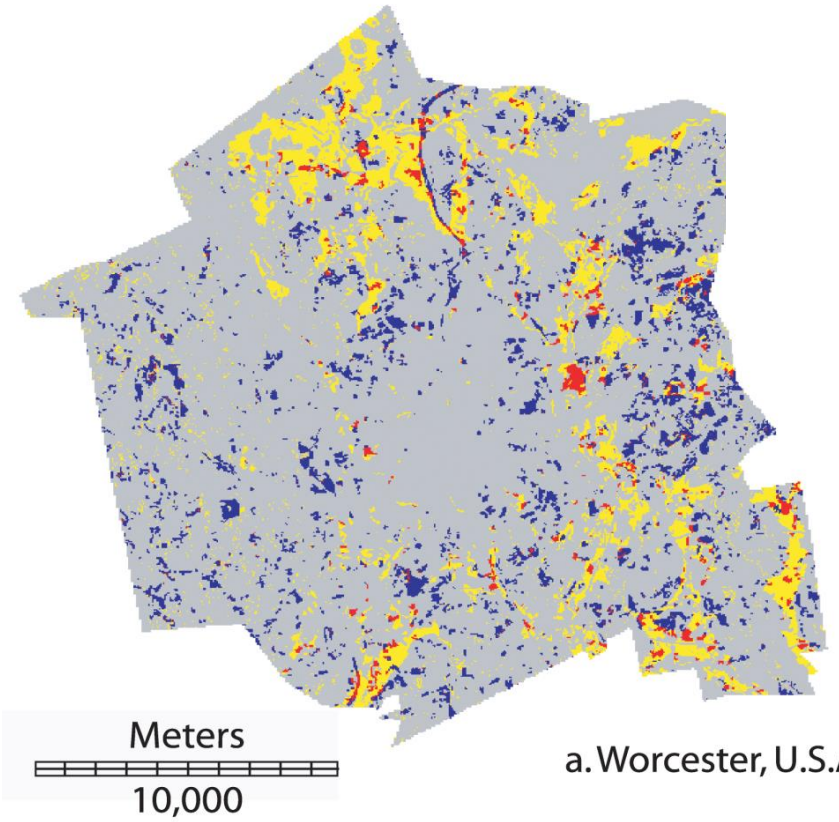
a. Worcester, U.S.

There is more error than correctly predicted change.

Most of the error is due to predicting the wrong location by not more than 4 kilometers.

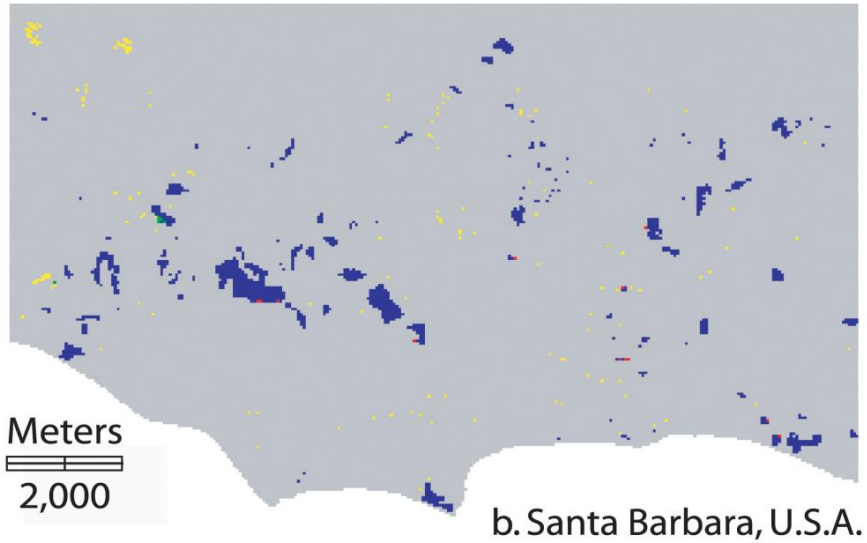
Misses	ERROR DUE TO OBSERVED CHANGE PREDICTED AS PERSISTENCE
Hits	CORRECT DUE TO OBSERVED CHANGE PREDICTED AS CHANGE
Wrong Hits	ERROR DUE TO OBSERVED CHANGE PREDICTED AS WRONG GAINING CATEGORY
False Alarms	ERROR DUE TO OBSERVED PERSISTENCE PREDICTED AS CHANGE
Correct Rejections	CORRECT DUE TO OBSERVED PERSISTENCE PREDICTED AS PERSISTENCE
	NOT CANDIDATE FOR TRANSITION
	OUT OF STUDY AREA

Geomod and Sleuth



a. Worcester, U.S.A.

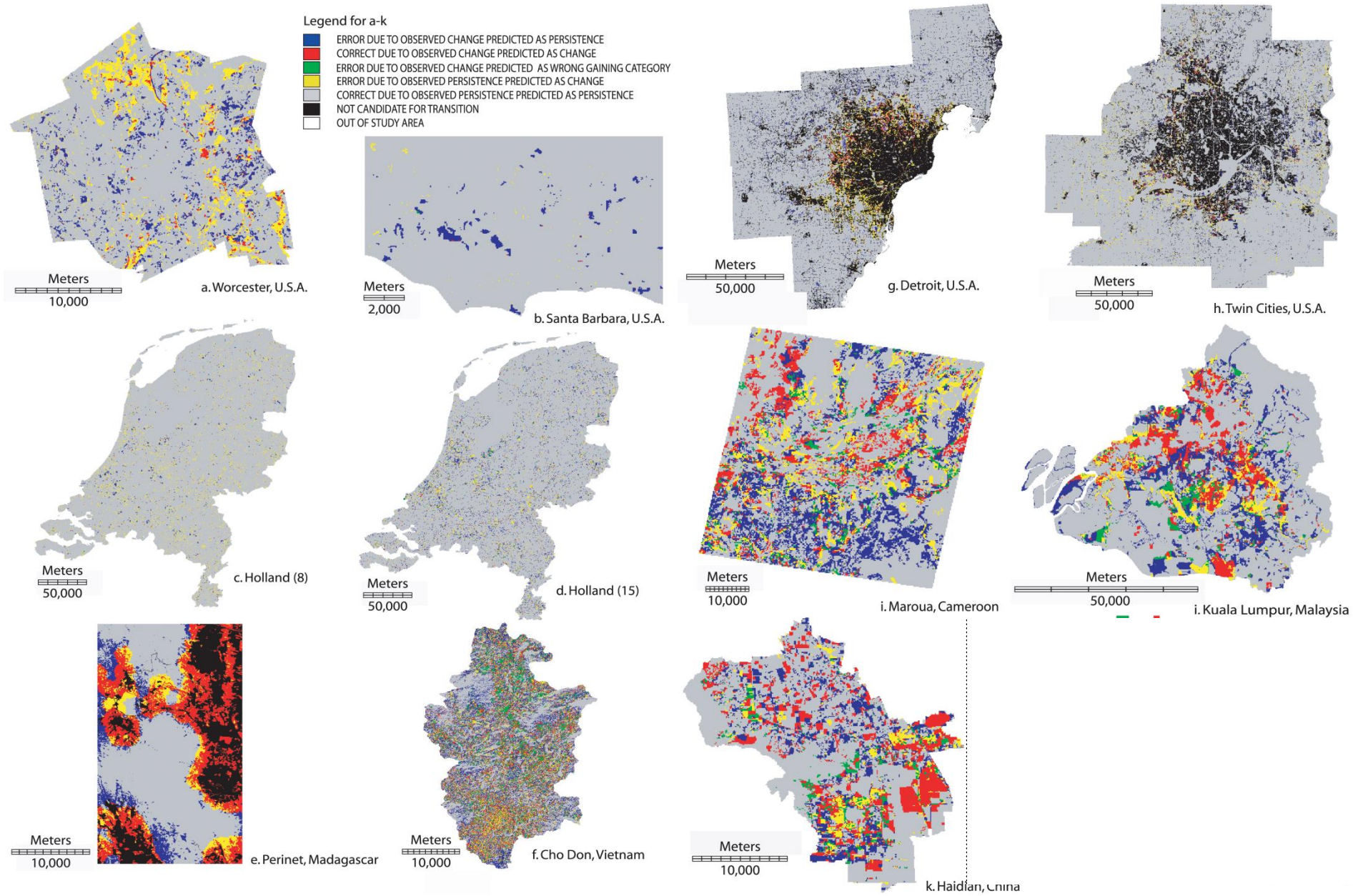
Most of the error is due to predicting less than the correct quantity of change.



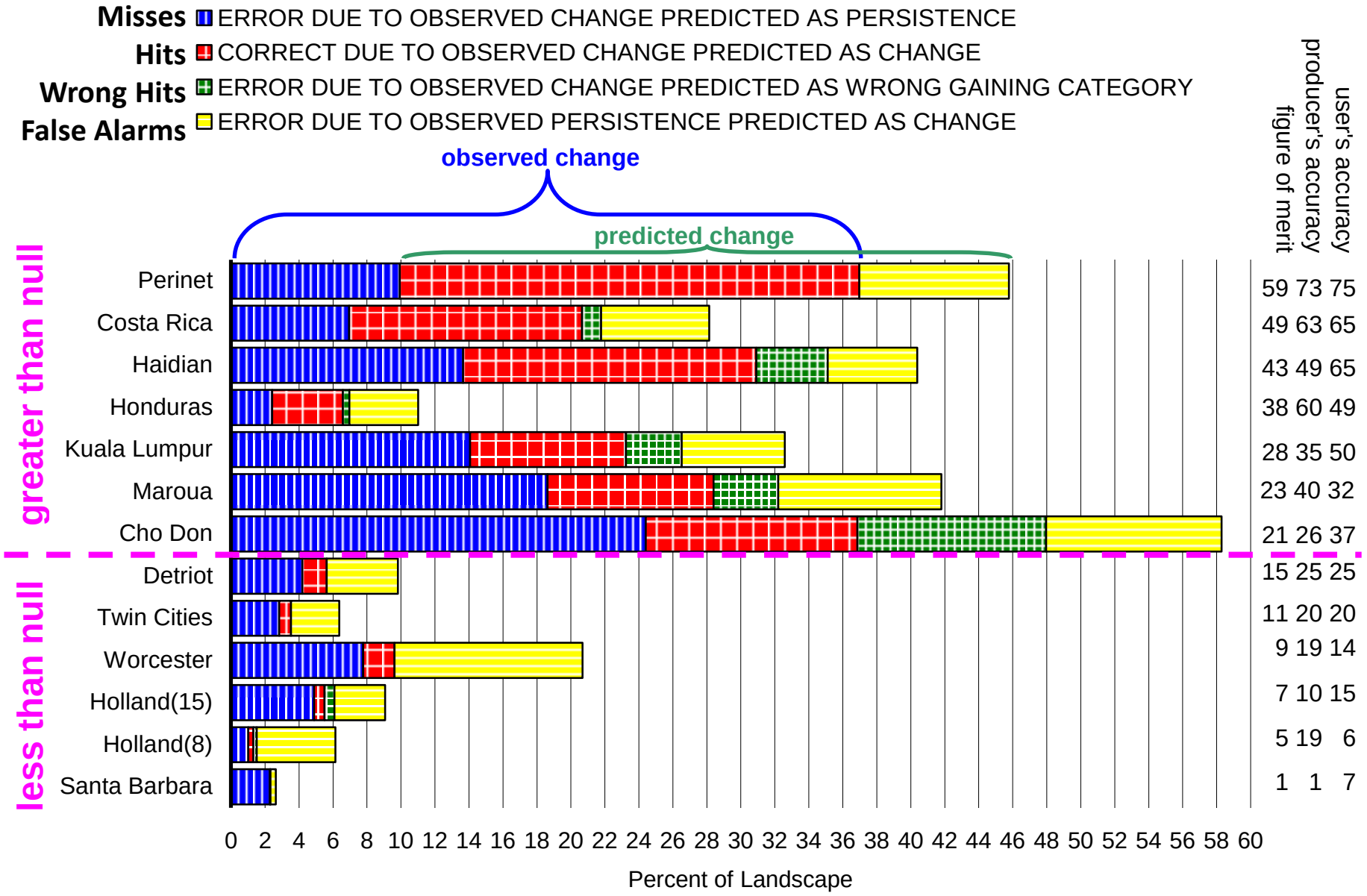
b. Santa Barbara, U.S.A.

Misses	ERROR DUE TO OBSERVED CHANGE PREDICTED AS PERSISTENCE
Hits	CORRECT DUE TO OBSERVED CHANGE PREDICTED AS CHANGE
Wrong Hits	ERROR DUE TO OBSERVED CHANGE PREDICTED AS WRONG GAINING CATEGORY
False Alarms	ERROR DUE TO OBSERVED PERSISTENCE PREDICTED AS CHANGE
Correct Rejections	CORRECT DUE TO OBSERVED PERSISTENCE PREDICTED AS PERSISTENCE
	NOT CANDIDATE FOR TRANSITION
	OUT OF STUDY AREA

Which of these 11 cases show more hits than errors and why?



12 of 13 cases have more error than hits at the detailed resolution. “Less than null” means percent correct is less than a prediction of no change.



Response from non-modelers

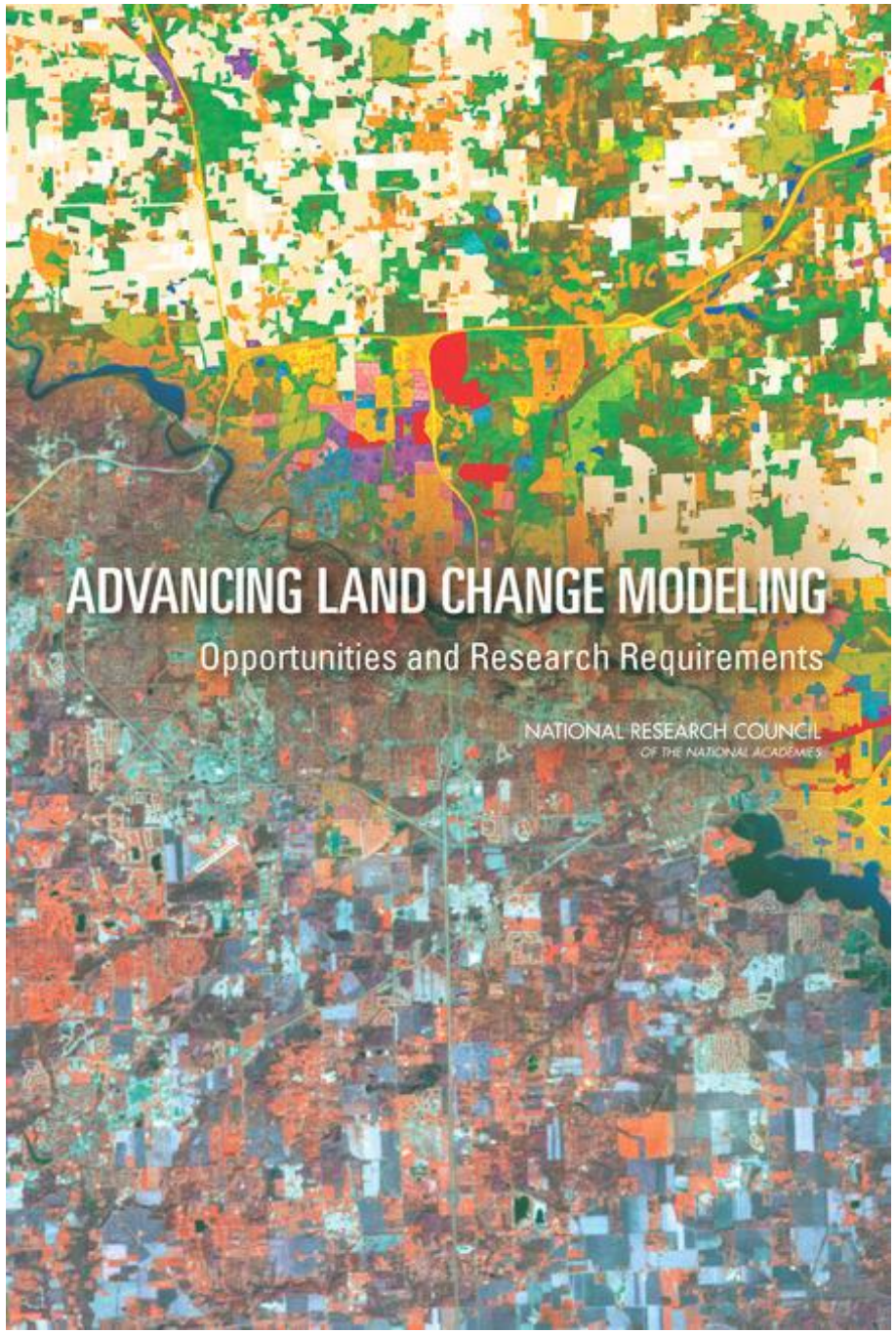
"Your colleagues must hate you!"

Response from modelers

“Thank you for exposing this,

because now I can publish any results!”

Brown, Daniel G, Lawrence E Band, Kathleen O Green, Elena G Irwin, Atul Jain, Eric F Lambin, Robert G Pontius Jr, Karen C Seto, B L Turner II, Peter H Verburg. 2013. ***Advancing Land Change Modeling: Opportunities and Research Requirements***. The National Academies Press: Washington DC. 145 pages.



Brown et al. (2013) Opportunities to improve impact, integration, and evaluation of land change models

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Environmental Sustainability

Opportunities to improve impact, integration, and evaluation of land change models

Daniel G Brown[1], Peter H Verburg[2], Robert Gilmore Pontius Jr[3] and Mark D Lange[4]

Land change modeling supports analyses, assessments, and decisions concerning land management by providing a platform for both encoding mechanisms of land-change processes and making projections of future land-cover and land-use patterns. Approaches have ranged from pattern-based methods, such as machine learning models, to structural or process-based methods, such as economic or agent-based models. Selection of the appropriate modeling approach for a given scientific or decision making purpose is essential. Additionally, we argue that more needs to be done to develop and disseminate methods for evaluating land-change models

reviews was how to represent processes of human decision making in these models as a mechanism by which land changes are made. Work over the last decade has focused on representing human decision making, coupling between human and environmental systems, and addressing questions about environmental sustainability challenges through model coupling [7]. Thus, more recent reviews have provided more complete coverage of models that integrate across human and natural systems and explicitly represent how human actors behave in these systems [2,8,9], which are advances that

Brown et al. (2013) conclude that scientists should:

1. align model choices with modeling goals.
2. integrate Land Change Models with observational data, across scales, and across positive and normative modeling approaches.
3. improve and disseminate use of model evaluation approaches.

Brown et al. (2013) did not conclude that scientists should:

1. strive to make models that models that predict more accurately.
2. use data and models that are more detailed and complex.
3. apply the most expensive computer technology.

But this is what Pontius sees on most research agendas.

Stories of doctoral students Orsi Varga, Claudia Viana, Aiyin Zhang, and Thomas Bilintoh

These doctoral students have read the literature.

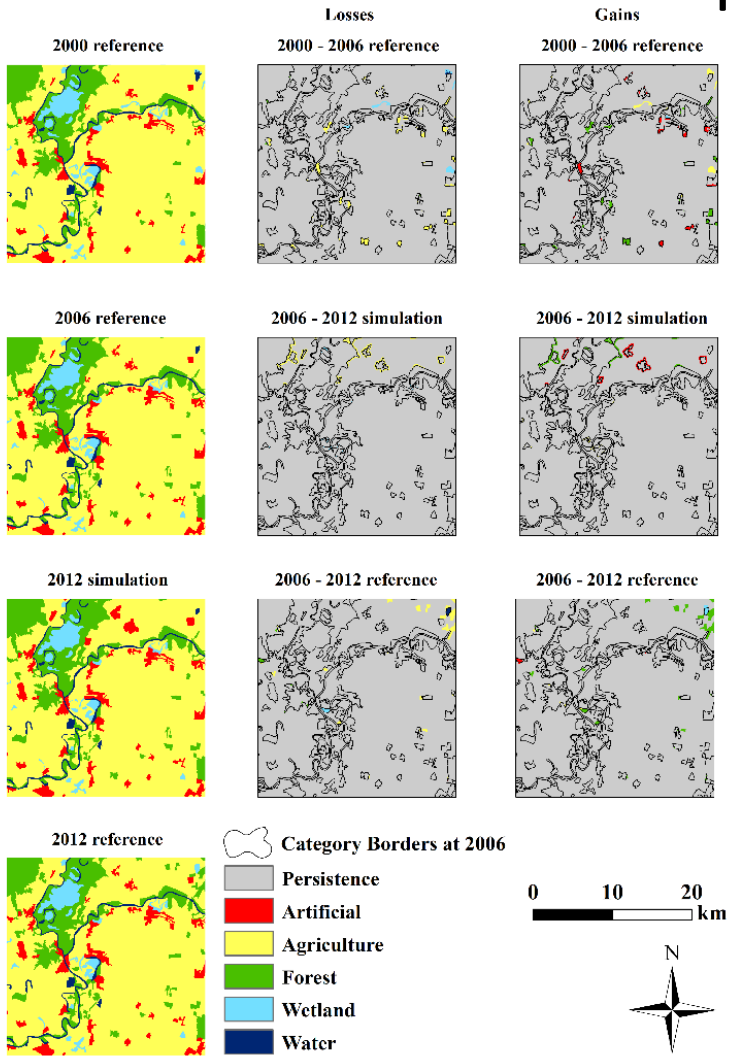
They worked with Pontius to satisfy university requirements, to maintain their scientific integrity, to advance their careers, and to enjoy doing it.

Orsi Varga came to Pontius to kill kappa, which Pontius said he already did.

Varga told Pontius that he is wrong, because she sees people still using kappa.

Pontius realized that Varga is correct.

Cellular Automata-Markov example from Orsi Varga



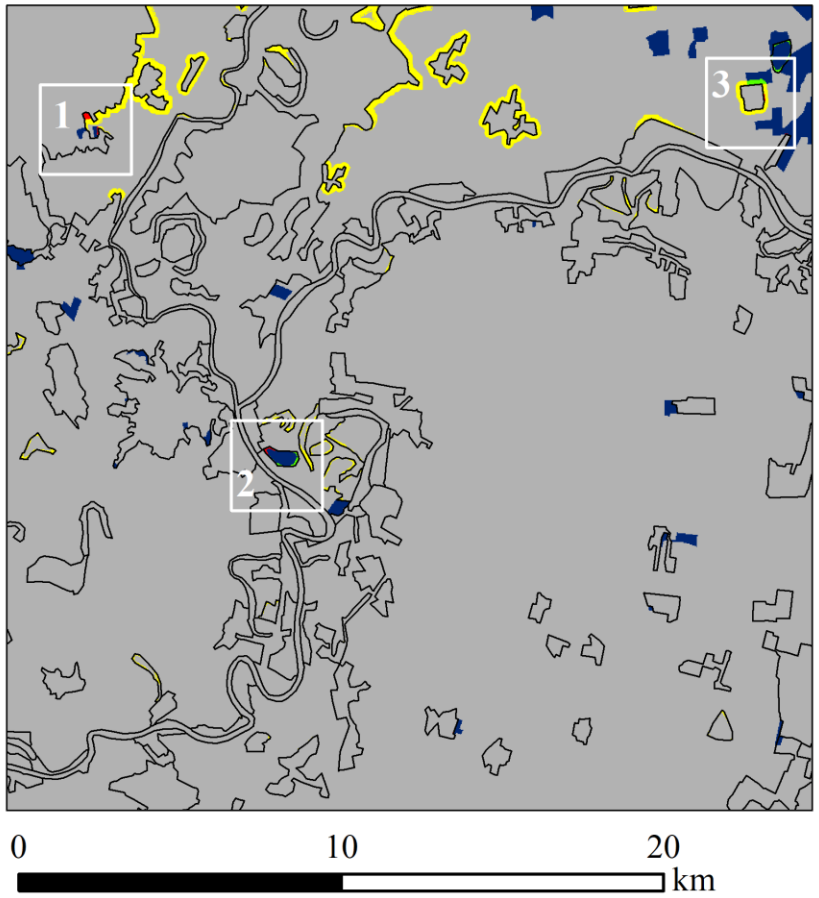
Calibration Interval is 2000-2006.

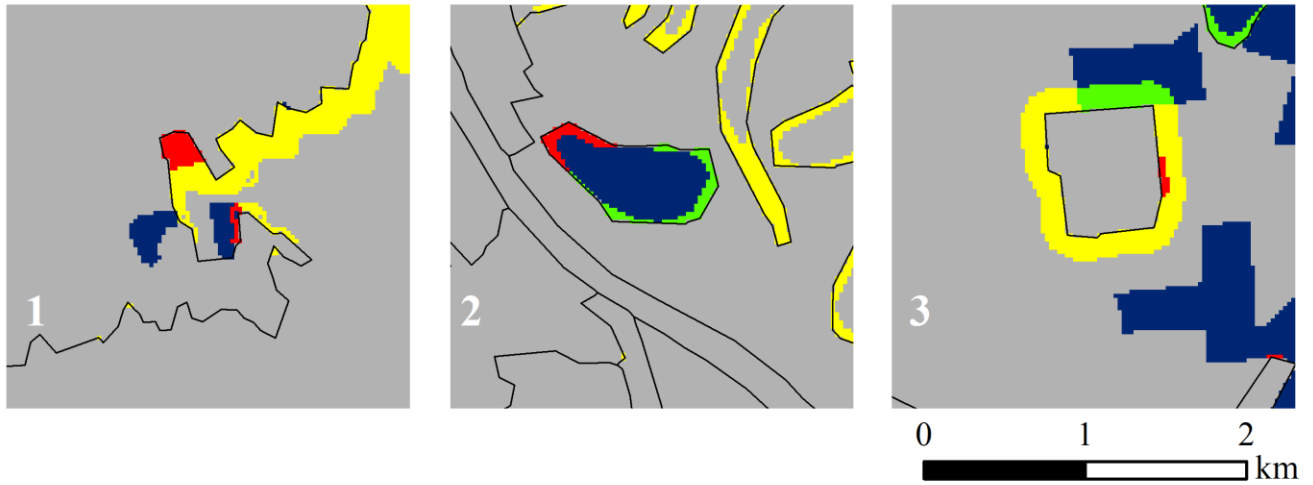
Simulation Interval is 2006-2012.

Validation Interval is 2006-2012.

Varga, Orsolya Gyöngyi, Robert Gilmore Pontius Jr, Sudhir Kumar Singh, Szilárd Szabó. 2019. Intensity Analysis and the Figure of Merit's Components for assessment of a Cellular Automata - Markov simulation model. Ecological Indicators 101: 933-942.

Zoom into see Misses, Hits, Wrong Hits and False Alarms





Category Borders at 2006

Correct Rejections = reference persistence simulated correctly (97.41%)

False Alarms = reference persistence simulated as change (1.47%)

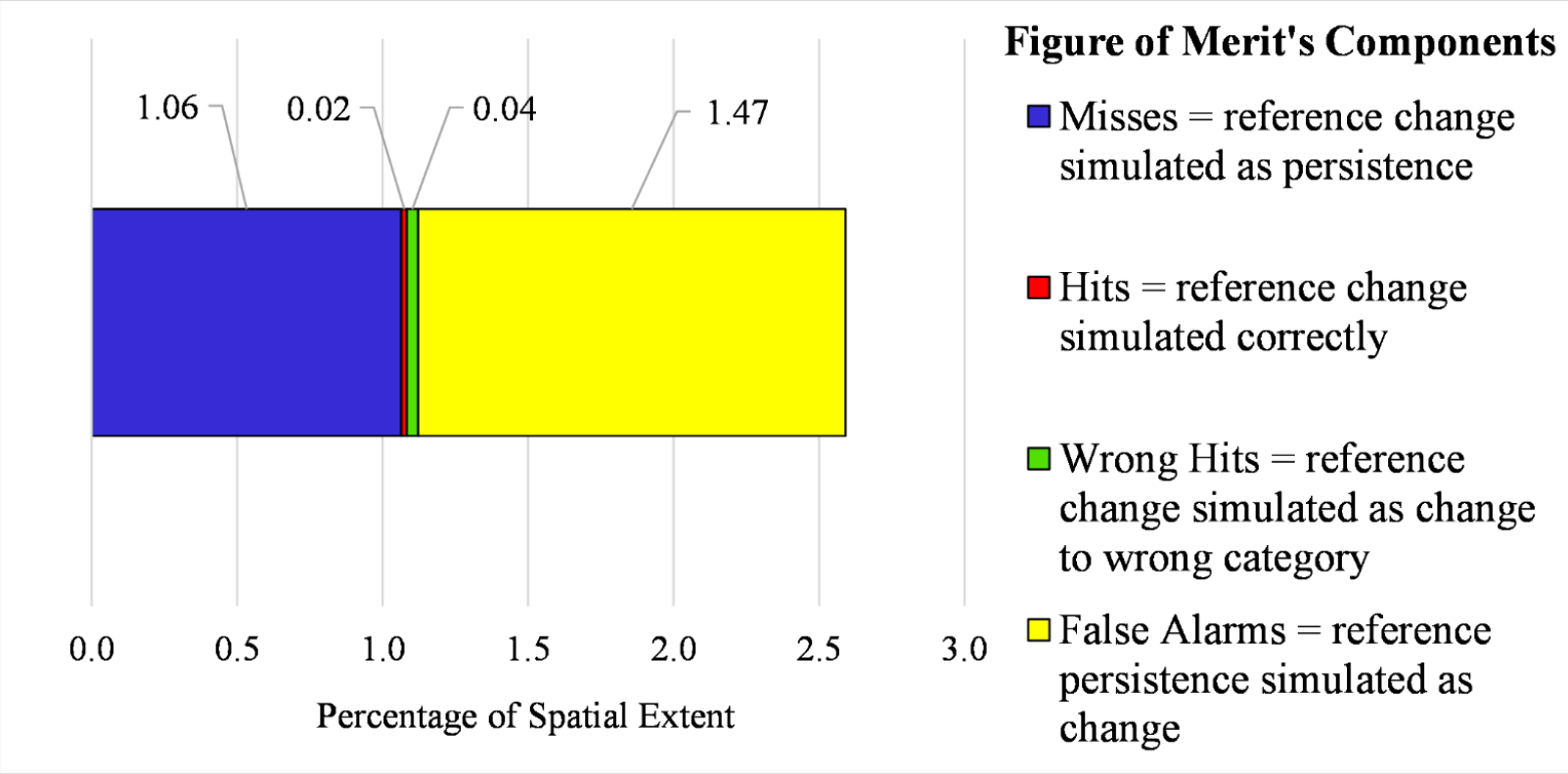
Wrong Hits = reference change simulated as change to wrong category (0.04%)

Hits = reference change simulated correctly (0.02%)

Misses = reference change simulated as persistence (1.06%)

Varga, Orsolya Gyöngyi, Robert Gilmore Pontius Jr, Sudhir Kumar Singh, Szilárd Szabó. 2019. Intensity Analysis and the Figure of Merit's Components for assessment of a Cellular Automata - Markov simulation model. Ecological Indicators 101: 933-942.

Misses < False Alarms because reference change decelerates from the calibration to validation intervals, which produces quantity disagreement. The disagreement's source is a non-stationary landscape, not the model.



Varga, Orsolya Gyöngyi, Robert Gilmore Pontius Jr, Sudhir Kumar Singh, Szilárd Szabó. 2019. Intensity Analysis and the Figure of Merit's Components for assessment of a Cellular Automata - Markov simulation model. Ecological Indicators 101: 933-942.

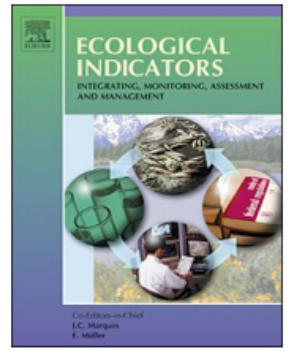
Varga gets a publication with her professors as co-authors.

Ecological Indicators 101 (2019) 933–942

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind



Original Articles

Intensity Analysis and the Figure of Merit's components for assessment of a Cellular Automata – Markov simulation model

Orsolya Gyöngyi Varga[a,*], Robert Gilmore Pontius Jr.[b], Sudhir Kumar Singh[c,1], Szilárd Szabó[a]

[a] *Department of Physical Geography and Geoinformatics, Faculty of Science and Technology, University of Debrecen, Debrecen, Hungary*
[b] *School of Geography, Clark University, Worcester, MA 01610, USA*
[c] *K. Banerjee Centre of Atmospheric & Ocean Studies, IIDS, Nehru Science Centre, University of Allahabad, Prayagraj 211002, Allahabad, Uttar Pradesh, India*

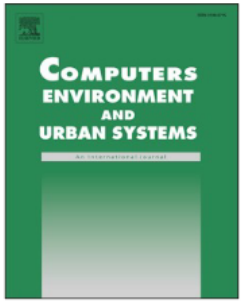
Pontius spent several rounds reviewing this paper, which now concludes that a model does not have an intrinsic accuracy.

Computers, Environment and Urban Systems 104 (2023) 102004

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



Check for updates

A user-friendly assessment of six commonly used urban growth models

Yuzhi Zhang [a], Mei-Po Kwan [b], Jun Yang [a,*]

[a] Department of Earth System Science, Institute for Global Change Studies, Ministry of Education Ecological Field Station for East Asian Migratory Birds, Tsinghua University, Beijing 100084, China

[b] Department of Geography and Resource Management and Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Shatin, Hong Kong, China

ARTICLE INFO

Keywords:
Urban growth model

ABSTRACT

An accurate grasp of urban expansion patterns is conducive to efficient urban management and planning. Various urban growth models have been developed to meet this need in the last two decades. As more models become

Claudia Viana read that the CA-Markov is helpful for managing landscapes.
Viana used CA-Markov then asked Pontius to proofread her manuscript.
We wrote a paper to recommend four criteria to define a good model.

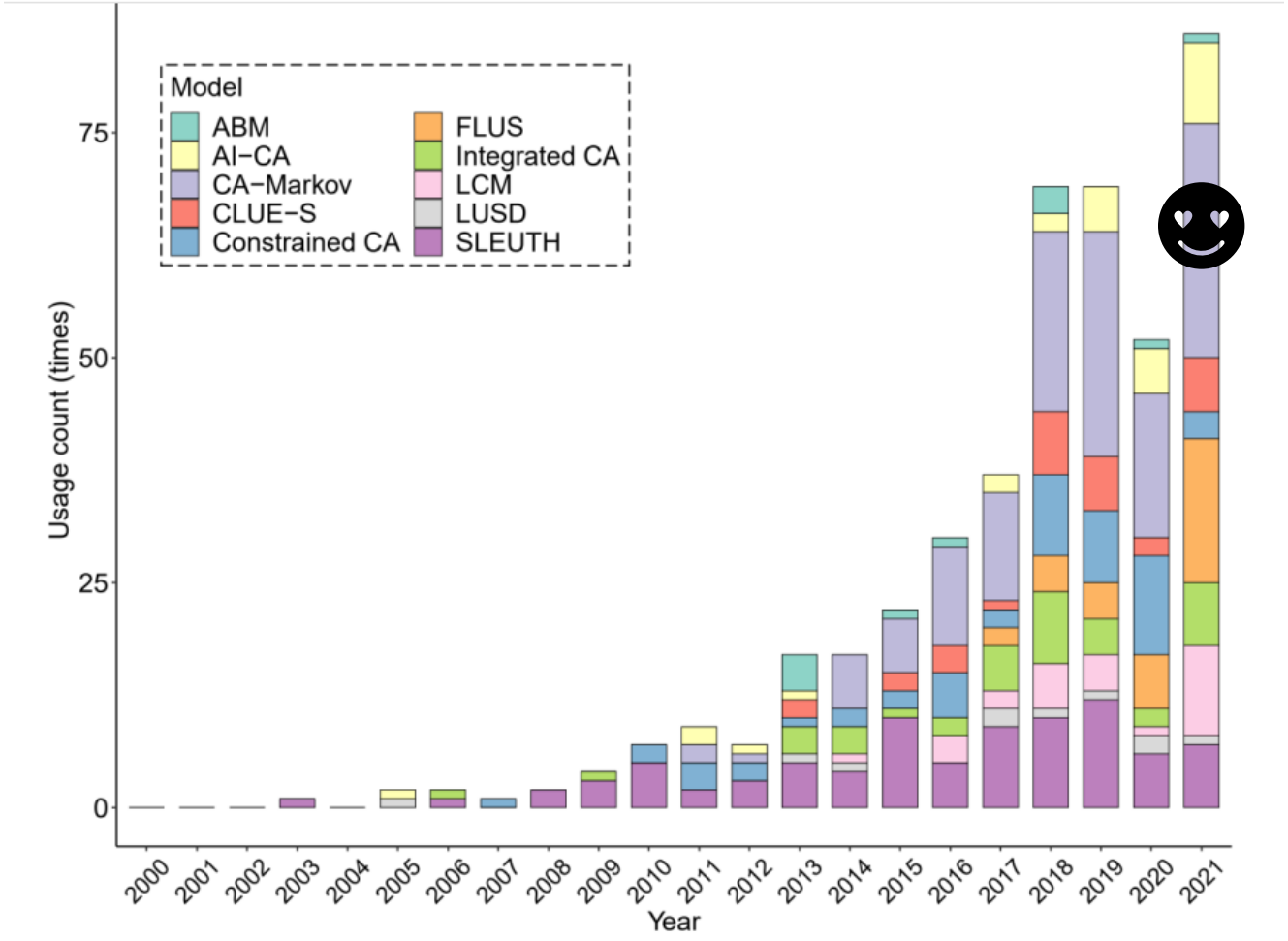
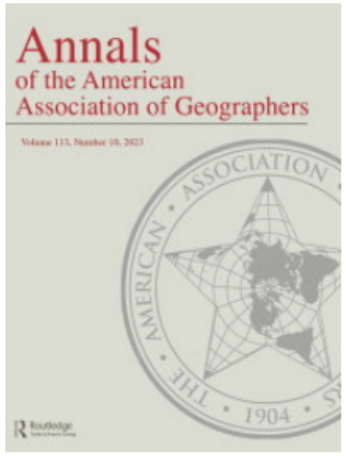


Figure 1 Usage counts of the top ten urban growth models between 2000 and 2021.

Zhang, Kwan, and Yang. 2023.
A user-friendly assessment of six commonly used urban growth models.
Computers, Environment and Urban Systems.

Viana gets an excellent publication.



Taylor & Francis
Taylor & Francis Group

Annals of the American Association of Geographers

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/raag21

Four Fundamental Questions to Evaluate Land Change Models with an Illustration of a Cellular Automata–Markov Model

Cláudia M. Viana, Robert Gilmore Pontius Jr. & Jorge Rocha

To cite this article: Cláudia M. Viana, Robert Gilmore Pontius Jr. & Jorge Rocha (2023) Four Fundamental Questions to Evaluate Land Change Models with an Illustration of a Cellular Automata–Markov Model, Annals of the American Association of Geographers, 113:10, 2497-2511, DOI: 10.1080/24694452.2023.2232435

To link to this article: https://doi.org/10.1080/24694452.2023.2232435

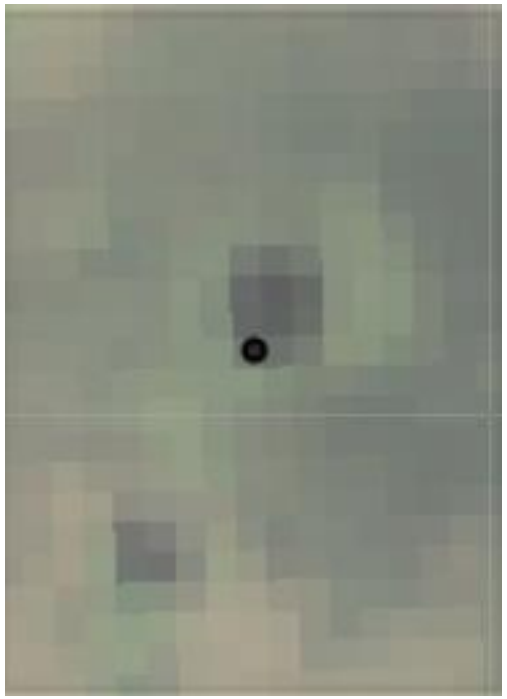
Four fundamental questions to evaluate land change models with an illustration of a Cellular Automata – Markov model

Our manuscript raises four fundamental questions to help users decide whether to use any model: 1) Can the user understand the model? 2) Can the audience understand the model? 3) Can the user control the model? and 4) Does the model address the goals of the specific application?

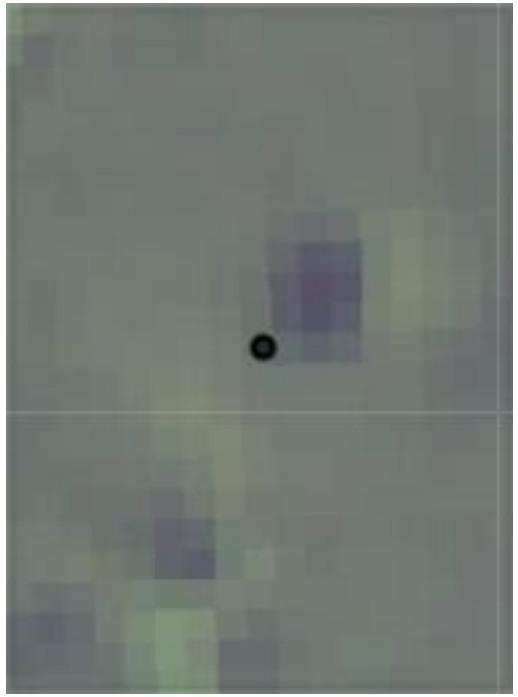
Results show that the model behaves in ways that the model's documentation fails to describe. The model's behavior is likely to cause users to misinterpret the validation metrics.

Aiyin Zhang proposed to make a time series of maps that show change of water.
Her team looked at the images to ask “Is there change of water at the sample point?

Time 1



Time 2



Time 3

Aiyin Zhang leads a team of students at Clark University.

The images are inconsistently georegistered.
Various interpreters give different assessments.
Interpreters are uncertain, which means the reference data are unreliable.

American Association of Geographers awards Aiyin first place.

Dr. Robert Gilmore Pontius (PIE Investigator, left) and Aiyin Zhang (PIE Graduate Student, right) proudly posing in front of Aiyin's first-place award certificate.

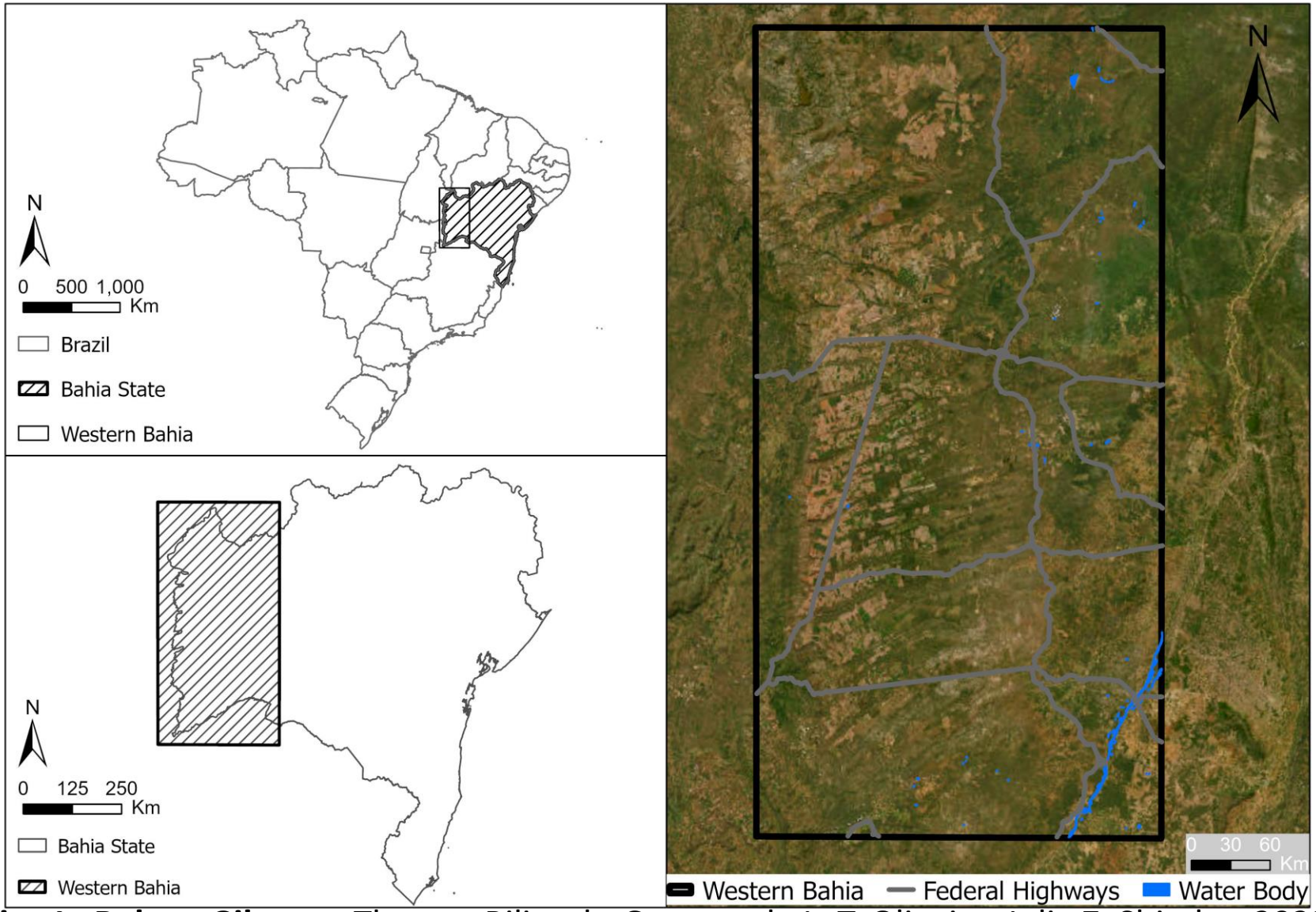
NATIONAL SCIENCE FOUNDATION
LTER NETWORK
LONG TERM ECOLOGICAL RESEARCH

Aiyin Zhang won first place in the Remote Sensing Specialty Group's Competition with her presentation titled "Data quality assessment of loss and gain of a land category during a time series".
Aiyin's results reveal the uncertainties of reference data and the misleading nature of popular accuracy metrics.

Pontius, Oliveira and Shimbo need graduate students to study irrigation in Brazil as an adaption to climate change.

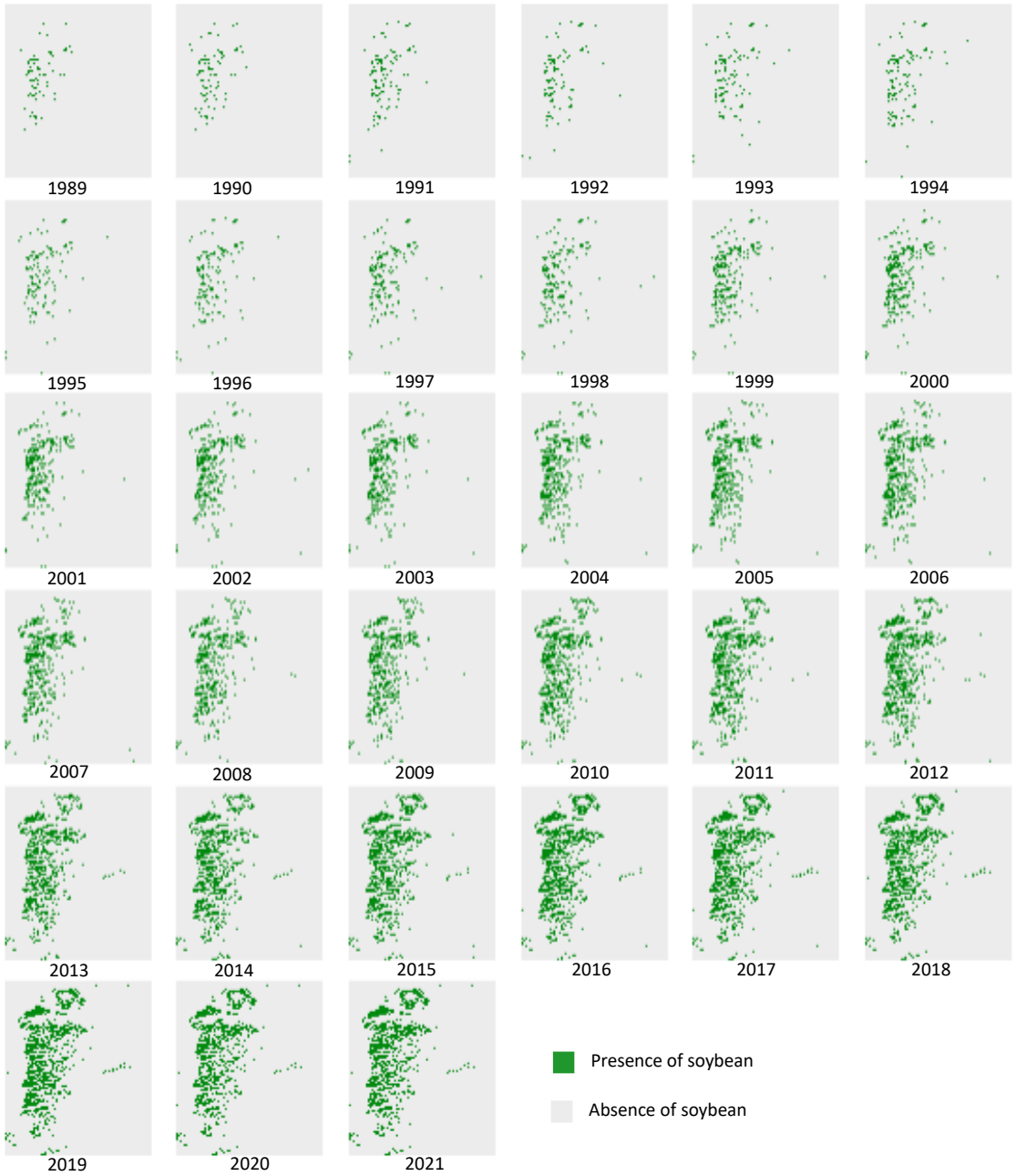
We use free data from MapBiomas.

Western Bahia is a hotspot for soybean cultivation.



Pontius Jr, Robert Gilmore, Thomas Bilintoh, Gustavo de L. T. Oliveira, Julia Z. Shimbo. 2023. TRAJECTORIES OF LOSSES AND GAINS OF SOYBEAN CULTIVATION DURING MULTIPLE TIME INTERVALS IN WESTERN BAHIA, BRAZIL. Space Week Nordeste. Fortaleza, Brazil.

Maps show soybean in western Bahia at 33 years forming 32 time intervals.



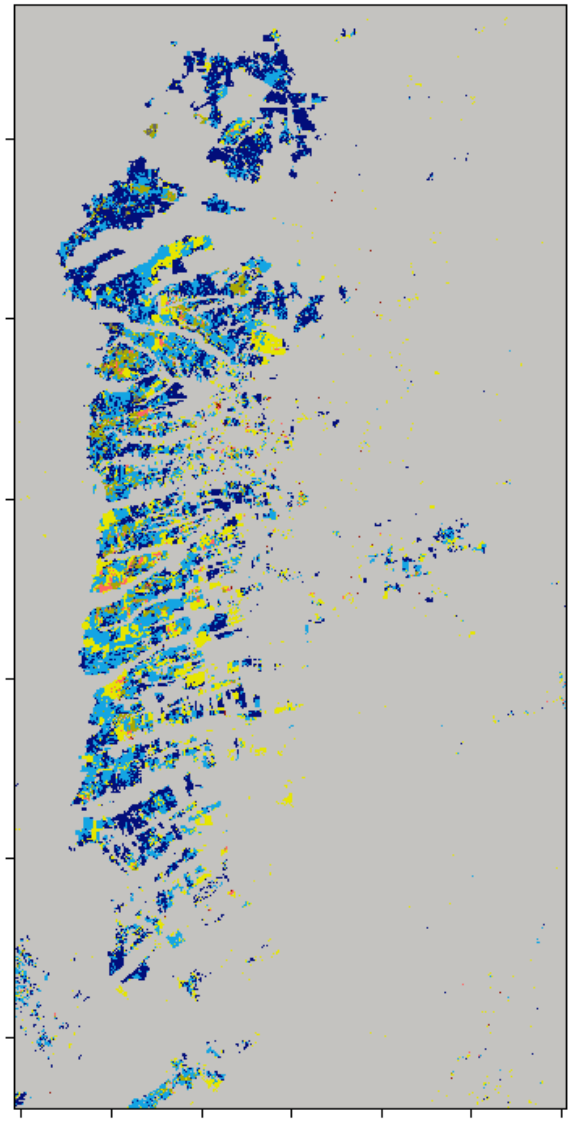
Each pixel has more than 8 billion possible combinations of presence or absence of soybean across the 33 years.

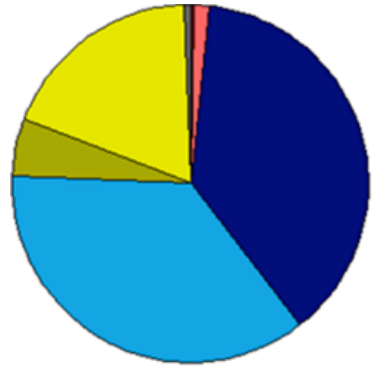
$2^{33} = 8.6 \text{ X } 10^{9}$

The challenge is to design graphics so that we can see important patterns and ignore relatively less important patterns.

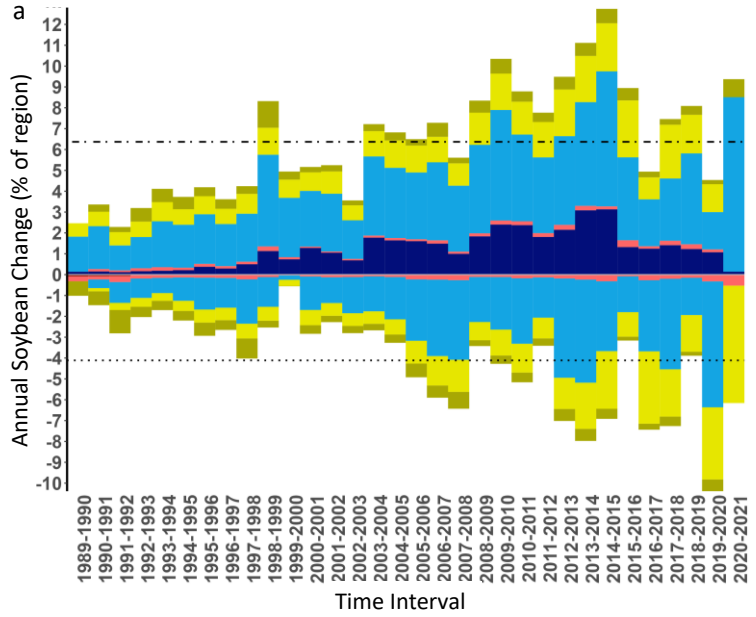
One map shows eight trajectories during 32 time intervals.

Use code at the GitHub site of Thomas Bilintoh.





Most of the change is Alternation.



Loss Without Alternation

Loss With Alternation

Gain Without Alternation

Gain With Alternation

All Alternation Loss First

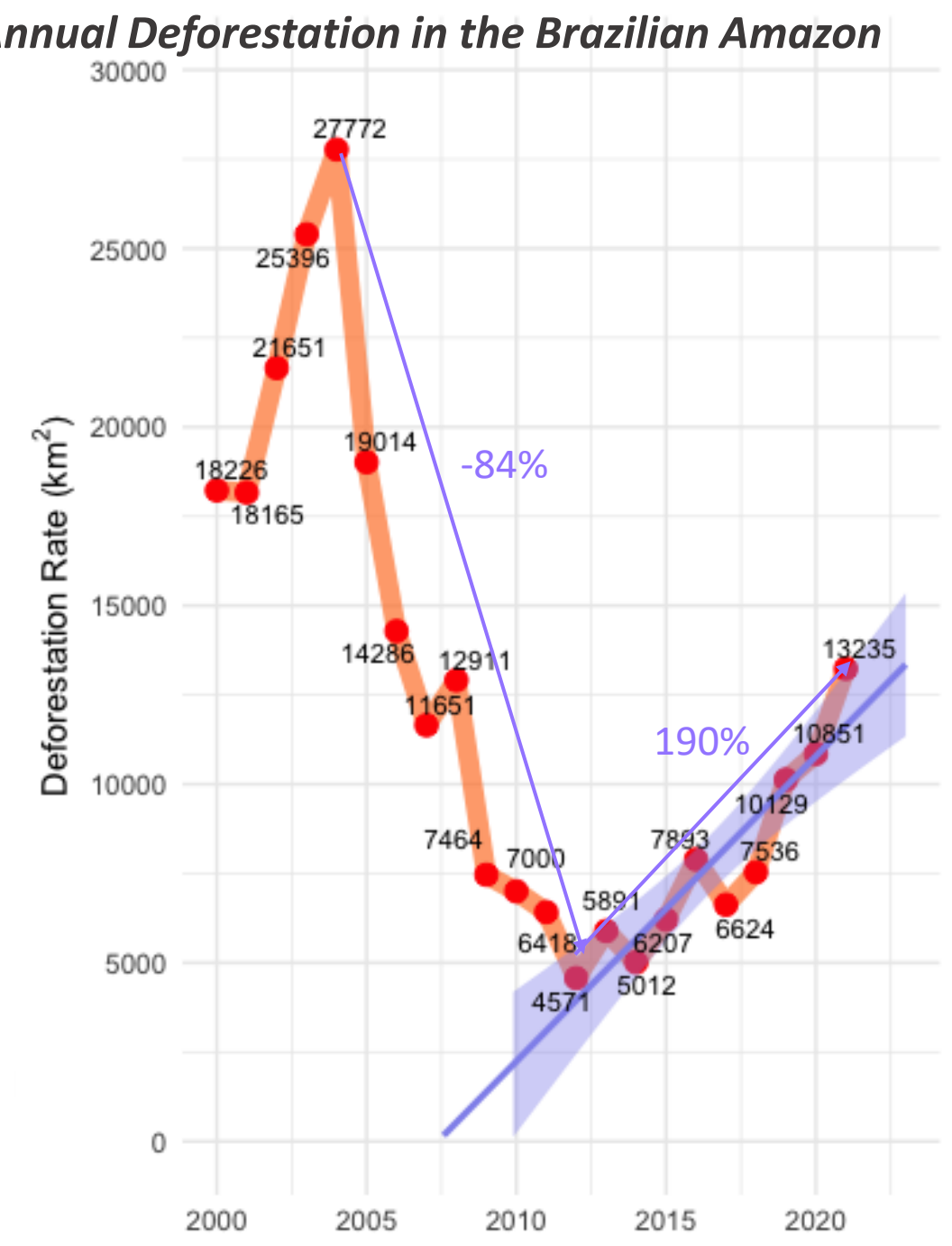
All Alternation Gain First

Stable Presence

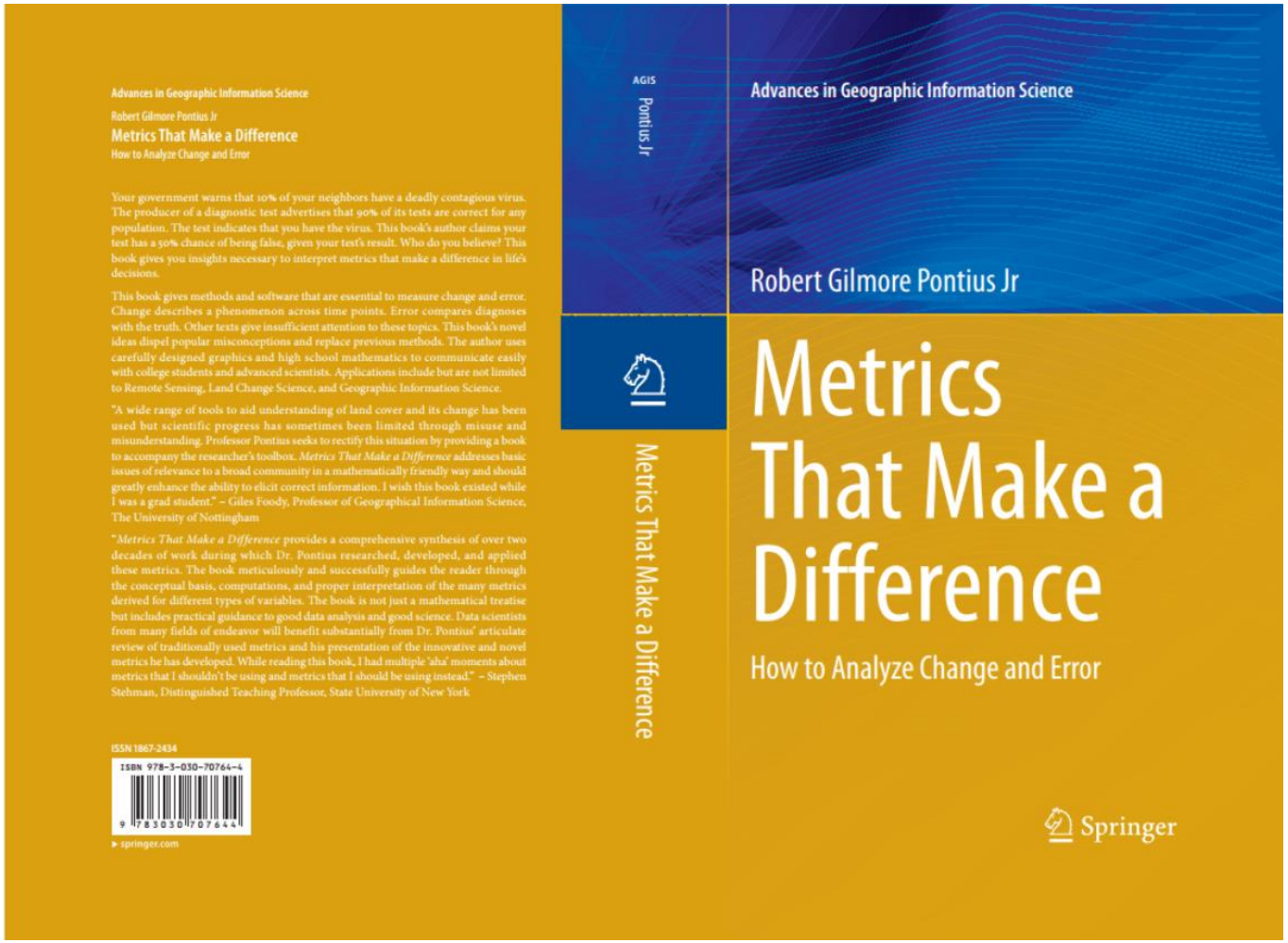
Stable Absence

How should we evaluate a prediction model when the pattern is non-stationary?

Pontius is working with Verra, TerraCarbon and Clark Labs on modeling for carbon offset projects to Reduce Emissions from Deforestation and forest Degradation (REDD).



Read this book available at www.clarku.edu/~rpontius starting with the chapter entitled "Commandments to avoid deadly sins"



Conclusions and Recommendations

1. Our profession has enormous problems, many of which are social problems related to our motivations, which respond to problematic incentive systems.
2. You can advance your career, keep your dignity, and enjoy your work, but you must be willing to report what you see.
3. Listen to students such as Orsi Varga, Claudia Viana, Aiyin Zhang, and Thomas Bilintoh.
4. Read the book Metrics That Make a Difference starting with the chapter entitled “Commandments to Avoid Deadly Sins”.