Advances in Geographic Information Science

Robert Gilmore Pontius Jr

Metrics That Make a Difference

How to Analyze Change and Error

Springer

Advances in Geographic Information Science

The series aims to: present current and emerging innovations in GIScience, describe new and robust GIScience methods for use in transdisciplinary problem solving and decision making contexts, illustrate GIScience case studies for use in teaching and training situations, analyze GIScience tools at the forefront of scientific research, and examine the future of GIScience in an expanding knowledge-based economy. The scope of the series is broad and will encompass work from all subject disciplines that develop and use an explicitly spatial perspective for analysis and problem-solving.

Advances in Geographic Information Science.

More information about this series at https://link.springer.com/bookseries/7712

Robert Gilmore Pontius Jr

Metrics That Make a Difference

How to Analyze Change and Error

Springer

Robert Gilmore Pontius Jr
Graduate School of Geography
Clark University
Worcester, MA, USA

ISSN 1867-2434 ISSN 1867-2442 (electronic)
Advances in Geographic Information Science
ISBN 978-3-030-70764-4 ISBN 978-3-030-70765-1 (eBook)
https://doi.org/10.1007/978-3-030-70765-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my son Nicholas and daughter Olivia who generate my guiding light

Photo by Nicholas K Pontius

Foreword

This book has been at least 20 years in the making. Land change science was a reasonably new field in 2000, with comparison of remote sensing observations of land use and cover over time as one of its core tools. While geographers, remote sensing scientists, and statisticians had developed tools for comparisons of maps and other forms of categorical variables by that time, there was a reasonable paucity of tools for fully understanding and quantifying these changes in ways that help unravel the key processes of interest to land change scientists, like urbanization, land degradation, rotational agriculture and forestry, and land-use intensification. The kappa statistic, originally developed for comparison of categorical variables in educational research, had been adapted as the key tool for making quantifying both error in remote sensing classifications and change between two land-cover maps, but it was a reasonably blunt tool for the latter purpose especially. Seeing the need for better tools to understand how land covers are changing and interpreting those changes in terms of processes of interest to domain scientists, Professor Pontius set out to construct a rich mathematical scaffolding that exploits the change matrix, a pairwise accounting of the amounts of land covers in each category at two different times, and provides tools for researchers to better conduct their work.

The results of his work, described in many papers and now summarized smartly in this volume, have informed and enabled hundreds of land change studies, and his generosity in making his algorithms available to others through an easy-to-use Excel spreadsheet and numerous training sessions, have contributed substantially to that impact. His efforts to deconstruct observed changes into mathematical descriptions of component parts of quantity, exchange, and shift, and to build up new summary statistics like total operating characteristic and intensity analysis have brought both fresh thinking to land change science and provided a general framework for map comparisons across a broader range of applications in geographic information science. While I understand that it was never his intention, these efforts at rethinking and recalculating categorical map comparisons led Professor Pontius to declare “Death to Kappa” in his most highly cited publication from 2011. The title along with that iconic paper signaled the birth of a whole new generation of quantitative map comparison statistics that will continue to serve, thanks also to the introduction of this volume, in advancing land change and other sciences now and on into the future.

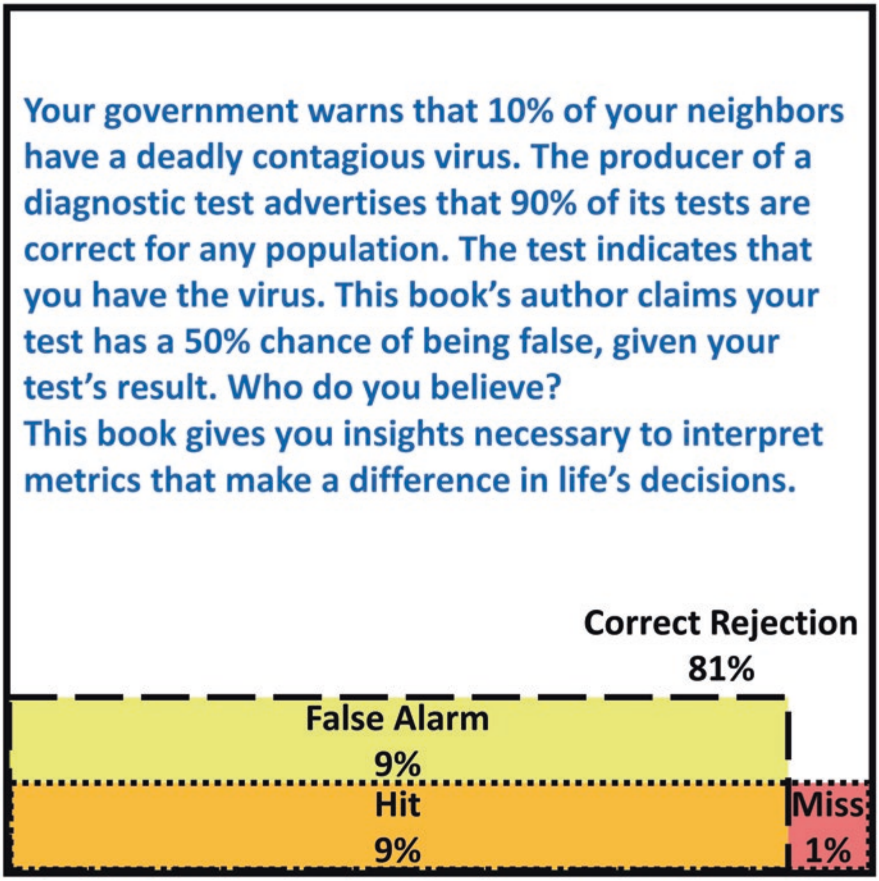
Daniel G. Brown
Professor and Director, School of Environmental and Forest Sciences,
Washington University,
Seattle, WA, USA

Preface

Let us begin with a brainteaser so you can get an idea of the concepts you will learn in this book. Take a moment to ponder the riddle before you read the next paragraph. Your government warns that 10% of your neighbors have a deadly contagious virus. The producer of a diagnostic test advertises that 90% of their tests are correct for any population. The test indicates that you have the virus. This book's author claims your test has a 50% chance of being false, given your test is positive. Who do you believe? This book gives you insights necessary to interpret metrics that make a difference in life's decisions.

The solution is that the government, the producer, and the professor are giving consistent information, yet some metrics are more helpful than other metrics. The producer's advertisement that 90% of all tests are correct is partially helpful but potentially misleading for your purpose. You need a metric that makes a difference to your decisions. You need to know the probability of having the virus, given that the test diagnosed the Presence of the virus. If you had read this book, then you would have likely visualized the rectangular Venn diagram below, where the bounding square represents all tests (Fig. 1). The Venn diagram has two sets drawn to scale for our example. The dotted boundary along the bottom outlines the set of true Presence of the virus. The dashed boundary indicates the set of diagnosed Presence of the virus. The label Hit denotes the sets' intersection, which contains the True Positives. Half of the dashed set is in the dotted set. Your test diagnosed the Presence of the virus so you are in the dashed set; therefore, you have a 50% chance of truly having the virus.

This book offers metrics that make an important difference for interpretation, and warns you of metrics that do not. This book's intended audience ranges from undergraduate university students to senior scientists. Most of the mathematics in this book are addition, subtraction, multiplication, and division. Some of the later chapters use high-school-level concepts such as statistical regression. A major concept is a Venn diagram, which you probably have seen since your middle school math class. If you can understand a Venn diagram, then you already have a grasp of a major concept in most of this book. I write intentionally to communicate clearly with readers who might have math anxiety, while I suspect that many of the readers



Your government warns that 10% of your neighbors have a deadly contagious virus. The producer of a diagnostic test advertises that 90% of its tests are correct for any population. The test indicates that you have the virus. This book's author claims your test has a 50% chance of being false, given your test's result. Who do you believe?
This book gives you insights necessary to interpret metrics that make a difference in life's decisions.

Fig. 1 Sizes of Hits, Misses, False Alarms and Correct Rejections drawn to scale for brainteaser

enjoy math as I do. This book uses math to express concepts that are fundamental to science. If you find science valuable, then this book is for you.

I have generated the ideas in this book by thinking about its concepts for more than two decades. I have developed techniques to communicate the ideas while repeatedly teaching courses that I developed at Clark University in the United States. The ideas concern how to compare variable **X** with variable **Y**, where the observations form pairs (**X**,**Y**), where both **X** and **Y** show the same phenomenon. For example, one possible application is to assess diagnoses, where one variable describes the diagnosis and the other variable describes the truth. A second possible application is to compare two diagnoses, where **X** describes one diagnosis and **Y** describes another diagnosis, while the truth remains unknown. A third possible application is to characterize temporal change of a phenomenon, where **X** describes the start time and **Y** describes the end time. I give cases for several types of variables: binary, rank, categorical, interval, and vector. I am an applied statistician, thus I describe methods so readers can apply them to a variety of scientific subjects:

Biology, Computer Science, Chemistry, Engineering, Environmental Science, Management, Physics, Political Science, Psychology, Sociology, et cetera. My specialty is Geographic Information Science, thus the examples in this book relate to Geography. Scientists who analyze diagnostic errors, temporal changes, or other types of differences will benefit from this book.

I write this book because I offer something constructive to fix many of the problems I see repeatedly in my profession. I have seen the same types of problems in the hundreds of articles that I have reviewed for scientific journals. I see some of the same flaws in published literature and at scientific conferences. Frequently, I see a presentation of an elaborate method to generate a diagnosis or prediction, but then the assessment of the diagnosis or prediction applies methods that are popular, flawed and misleading. The methods frequently either make conceptual blunders or are unnecessarily complicated in ways that render the results uninterpretable. I cannot blame the authors, because authors typically follow methods that universities teach or that have become conventional in the profession due to unfortunate and dysfunctional aspects in the culture of scientists. I write this book to offer help. This book's methods are more straightforward, interpretable and helpful than many of the complicated and misleading methods that I see in the literature. Many metrics exist for the cases that this book considers. I have found that several popular measurements are unnecessarily complicated, frequently misinterpreted, and dangerously distracting. This book recommends the metrics to use and warns of metrics to avoid. I include methods that I have found to be relevant for many types of applications during my decades as a university professor, statistical consultant, and applied scientist.

I write to inspire hope. I hope this book guides others concerning how to present metrics to answer questions in ways that are clear and important for practical applications. Science is a discipline that requires focus, organization, and clarity; science is also an art that requires its practitioners to decide what details to ignore or to demote to lesser importance. This book's methods focus on the most fundamental issues, which one must understand before trying to interpret more subtle details.

A reviewer once described my work this way: *These methods are straightforward, thus any clear-thinking scientists should use them.* I was pleased with that comment because that is my goal. However, the reviewer apparently intended that comment as a negative criticism, which reflects a scientific culture that places value on complicated mathematics. I have found that if I focus on fundamental concepts, then the mathematics are simpler and thus easier to understand. I hope is that you use this book to clarify thoughts, to communicate results, to improve science, and to widen your audience.

Experienced scientists will find in this book several novel ideas that build from familiar fundamental concepts. First, this book's overall approach might be new for some readers because the book focuses on difference, whereas other popular literature focuses on agreement. I have found that metrics of difference are more effective than metrics of agreement at directing attention to the more important information. Differences can indicate errors, which are opportunities for improvement. Differences can indicate change, which is frequently the focus of temporal analysis.

Second, a fundamental concept is the contingency table for a categorical variable, while this book's relatively new concepts include three components of difference: Quantity, Exchange, and Shift. Third, some readers might be familiar with the Relative Operating Characteristic, while this book describes a more informative approach called the Total Operating Characteristic. Fourth, the chapter concerning multiple resolution analysis gives a method to address an issue that many scientists have encountered but have not known how to address, specifically how to distinguish minor Allocation errors from major Allocation errors. Fifth, the chapter that focuses on sampling gives a necessary procedure to convert from sample data to estimated population sizes, which some scientists fail to do. Sixth, the chapters concerning an interval variable gives concepts concerning linear regression, which exist in many software packages. The same chapter defines the difference components of Quantity and Allocation, which are fundamental but appear insufficiently in literature and software. Seventh, the chapter concerning Indices of Agreement describe metrics that are popular across fields or are common only in specialized fields. Eighth, I have rarely seen in practice methods to compare vector variables, which have both magnitude and direction. I include a chapter that offers a method to compare vector variables.

Some of the methods in this book are available in the GIS software TerrSet. Readers can learn more about TerrSet at https://clarklabs.org/. The PontiusMatrix42.xlsx file performs the calculations for the methods in Chaps. 1, 3, and 5. The Excel file is available for free at https://www.clarku.edu/faculty/rpontius/. My students have written software packages in the language R for some of the methods. Those R packages are available for free at https://cran.r-project.org/web/packages/. The diffeR package computes the concepts in Chaps. 1, 3, 4 and 7. The TOC package performs the analysis of Chap. 2. The intensity.analysis R package computes some of the metrics in Chap. 4. Videos concerning the techniques are at https://www.clarku.edu/faculty/rpontius/videos.html. I hope you experience as much enjoyment and insight in reading this book as I gained in writing it.

Worcester, MA, USA Robert Gilmore Pontius Jr

Acknowledgements

Most importantly, I thank my students, who helped me develop the concepts over the decades by working with me in courses, through theses, and via assistantships. They have contributed extremely helpful insights. I thank especially my doctoral advisees: Hao Chen, Safaa Aldwaik, Dan Runfola, Rahul Rakshit, Nicholas Giner, Alí Santacruz, Thomas Bilintoh, and Aiyin Zhang. I thank the students who wrote the 86 Masters and 15 Bachelors theses that I have advised. Some figures in Chap. 10 derive from the Honors thesis of Rebecca Chapman. The progress of science is an inter-generational process. Dr. Charles Hall was my doctoral advisor who inspired several approaches to science that contributed to how I matured as a scientist. One of his best pieces of advice is first to do the obvious straightforward analysis, then see what trouble it gets you into, then decide what to do next. A member of my doctoral committee was Dr. Steve Stehman, who contributed to my development in many ways, including by his lesson that statistical significance means something distinct from practical importance. Interactions with my professional colleagues inspired many of the ideas for this book. My colleagues at Clark University made an enjoyable and fruitful setting for intellectual creativity. The programmers at Clark Labs have incorporated some of the ideas in this book into the GIS software TerrSet. TerrSet's help section is the source of the data concerning sea surface temperature in Chap. 10. The Massachusetts Bureau of Geographic Information is the source of the data concerning land cover in Chaps. 6 and 7. Sam Khallaghi, Benoit Parmentier, Alí Santacruz, Kangping Si, and Amin Tayyebi wrote the diffeR and TOC packages in the language R. Priscilla Ahn and Priyanka Verma created the pontiPy software with the help of Claire Wang, Xiangyu Wen, and Zexing Zheng. The Edna Bailey Sussman Fund supplied funding to my students. The United States National Science Foundation funded two decades of my research through the Plum Island Ecosystems site of the Long Term Ecological Research network via a sequence of grants: DEB-972682, OCE-0423565, OCE-1026859, OCE-1058747, OCE-1238212, OCE-1637630.

Contents

List of Figures

List of Tables

About the Author

Robert Gilmore Pontius Jr My full name is Robert Gilmore Pontius Jr, while my friends call me Gil. I entered this world in 1963 in Pittsburgh, Pennsylvania, USA. I earned a Bachelor of Science in Mathematics at the University of Pittsburgh and then served two years in the Peace Corps teaching mathematics in Tanzania. I then completed a Master of Applied Statistics at The Ohio State University, before designing sampling frames as a Mathematical Statistician for the United States Department of Agriculture. My doctorate is in Environmental Science from the State University of New York's College of Environmental Science and Forestry in Syracuse. After one year as a Visiting Assistant Professor at Boston University, I gained two year's experience in a private environmental consulting company. Then I became extremely interested in juggling and made a living as a professional juggler for a year after winning the People's Choice Award of the International Jugglers Association. I considered a career in the performing arts, but I am glad I returned to science. I started at Clark University in 1998, where I have become a full professor in the Graduate School of Geography.

Chapter 1
Binary Variable Versus Binary Variable

Abstract This chapter gives methods to compare two variables, where both variables distinguish Presence from Absence of the same phenomenon. The analysis's foundation is a square contingency table that has four central entries: Hits, False Alarms, Misses, and Correct Rejections. Metrics express difference as the sum of two components: Quantity and Exchange. Some metrics express size in terms of the number of observations. Other metrics express intensities in terms of ratios of numbers of observations. Examples illustrate the importance of interpreting results in terms of the size and intensity of the components of Quantity and Exchange. Relevant software includes the PontiusMatrix42.xlsx spreadsheet (Pontius Jr 2020) available at www.clarku.edu/~rpontius and the diffeR package (Pontius Jr and Santacruz 2015) available at https://cran.r-project.org/web/packages/diffeR/index.html.

Keywords Binary · Exchange · PontiusMatrix · Quantity

1.1 Text

Binary is the name for the form of a variable that distinguishes between two possible states: Presence as opposed to Absence. Boolean is another name for this type of variable. Many practical applications use binary variables. For example, if each observation is a house, then variable **X** could indicate whether each house's smoke detector diagnoses the Presence or Absence of fire, while **Y** indicates the true Presence or Absence of fire. Both **X** and **Y** show the same phenomenon, which is the Presence or Absence of fire in houses. The analysis reveals the diagnostic ability of the detectors. As another example, if each observation is a place on a landscape, then **X** could indicate the Presence or Absence of wildfire at a start time, while **Y** indicates the Presence or Absence of wildfire at an end time. Both **X** and **Y** show the same phenomenon, which is the Presence or Absence of wildfire at various places. The analysis would analyze how wildfire changes through time across the landscape. This chapter gives methods to compare variable **X** to variable **Y** when both **X** and **Y** show the same phenomenon in the form of a binary variable.

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1_1

Figure 1.1 illustrates the concepts by analyzing ten observations for variables **X** and **Y**. The observations have the same sequence from left to right, thus form ten pairs. Three is the size of Presence in **X** while four is the size of Presence in **Y**. A contingency table summarizes how **X** compares to **Y**. The table is square in the respect that the table's number of rows equals its number of columns. The label for **X** is on the left and the label for **Y** is on the top of the table, which is a convention that this book and the profession follow. The convention is a good habit that keeps the concepts consistent thus clear. If an application analyzes a diagnosis, then **X** describes the diagnosis and **Y** describes the truth. If the application analyzes temporal change, then **X** describes the start time and **Y** describes the end time. The column at the far right and the row at the bottom are marginal sums. The sum at the far right gives the sizes of Presence and Absence in **X**, while the row at the bottom gives the sizes of Presence and Absence in **Y**. Extent is the name for the collection of all observations. The number in the table's bottom right is the size of the extent.

The square table has four central entries, where each entry gives the size for a combination of Presence versus Absence in **X** and **Y**. Hits are Presence for both **X** and **Y**. False Alarms are Presence for **X** and Absence for **Y**. Misses are Absence for **X** and Presence for **Y**. Correct Rejections are Absence for both **X** and **Y**. The selection of which variable is **X** and which variable is **Y** determines the definitions of Misses and False Alarms. If the application analyzes diagnostic power, then various professions use additional names for some of the four entries. Hits are also known as True Positives. False Alarms are also known as False Positives and Commission Errors. Misses are False Negatives and Omission Errors. Correct Rejections are True Negatives. If the application analyzes temporal change, then Hits are persistence of Presence. False Alarms are transitions from Presence to Absence, which are losses of Presence. Misses are transitions from Absence to Presence, which are gains of Presence. Correct Rejections are persistence of Absence.

Figure 1.1a gives a helpful way to envision the results in terms of sizes. The stacked bar is a horizontally oriented rectangular Venn diagram, which uses braces to show two sets. Presence in **Y** is the set at the left, while Presence in **X** is the set at the right. The intersection of the two sets is Hits. Misses are Presence in **Y** and not in **X**. False Alarms are Presence in **X** and not in **Y**. Figure 1.1a shows the size of Misses, Hits, and False Alarms as segments stacked from left to right, while the figure does not show explicitly Correct Rejections. Correct Rejections can be irrelevant and misleading for some applications. Correct Rejections are irrelevant when the extent is arbitrary. For example, suppose the application describes rare events, such as wildfires at two time points where the start time is **X** and the end time is **Y**. The extent could be where a scientist suspects any wildfire might exist on a landscape, thus various scientists might select various extents that contain all of the Presence observations for both **X** and **Y**. In this case, various extents would have the same size of Misses, Hits, and False Alarms, but larger extents would have more Correct Rejections. If various scientists report the size of Misses, Hits, and False Alarms, then all scientists would produce the same results regardless of extent. The Correct Rejections in a smaller extent would be fewer than the Correct Rejections in a larger extent. On the other hand, if the Correct Rejections are relevant to the

X	P	P	P	A	A	A	A	A	A	A
Y	P	A	A	P	P	P	A	A	A	A
Y versus X	H	F	F	M	M	M	C	C	C	C

		Y		
		Presence	**Absence**	**Sum**
X	**Presence**	Hits = 1	False Alarms = 2	**3**
	Absence	Misses = 3	Correct Rejections = 4	**7**
	Sum	**Abundance = 4**	**6**	**Extent Size = 10**

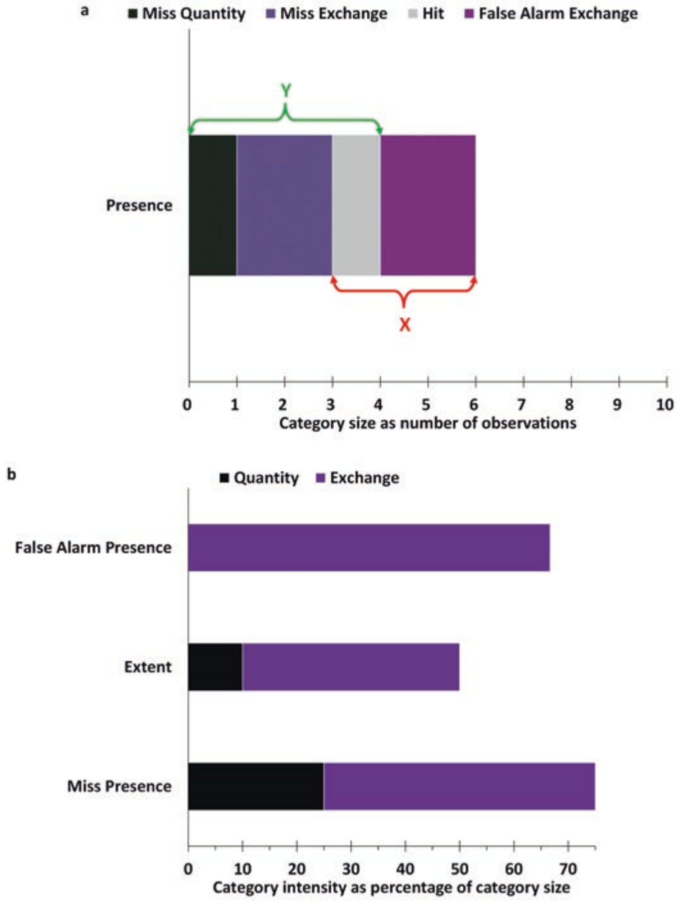


Fig. 1.1 Example where variables distinguish Presence (P) from Absence (A)

Table 1.1 Notation to compare two variables that show the same binary phenomenon

Notation	Meaning
C	Size of Correct Rejections, meaning Absence in both **X** and **Y**
F	Size of False Alarms, meaning Presence in **X** and Absence in **Y**
H	Size of Hits, meaning Presence in both **X** and **Y**
M	Size of Misses, meaning Absence in **X** and Presence in **Y**

analysis, then it might be informative for the horizontal axis in Fig. 1.1a to range from zero to the size of the extent, as Fig. 1.1a does.

Table 1.1 gives the mathematical notation for four numbers that this chapter's equations use. Those four numbers determine all entries in the contingency table. All the summary metrics derive from those four numbers. Equations 1.1, 1.2, 1.3, 1.4, 1.5 and 1.6 are the metrics that I have found most helpful to compute first. Equations 1.1, 1.2, 1.3, 1.4, 1.5 and 1.6 form a conceptual framework that the remainder of this book follows.

$$\text{False Alarm Quantity} = \text{MAXIMUM}\left(0, F - M\right) \tag{1.1}$$

$$\text{Miss Quantity} = \text{MAXIMUM}\left(0, M - F\right) \tag{1.2}$$

$$\text{False Alarm Exchange} = \text{Miss Exchange} = \text{MINIMUM}\left(F, M\right) \tag{1.3}$$

$$\text{Difference Quantity} = \text{False Alarm Quantity} + \text{Miss Quantity} = \left|M - F\right| \tag{1.4}$$

$$\text{Difference Exchange} = 2\,\text{MINIMUM}\left(F, M\right) \tag{1.5}$$

$$\text{Difference} = \text{Difference Quantity} + \text{Difference Exchange} = F + M \tag{1.6}$$

If the size of Presence in **X** is greater than the size of Presence in **Y**, then Eq. 1.1 gives a positive value for False Alarm Quantity; otherwise, Eq. 1.1 gives zero. If the size of Presence in **Y** is greater than the size of Presence in **X**, then Eq. 1.2 gives a positive value for Miss Quantity; otherwise, Eq. 1.2 gives zero. It is impossible for both Eqs. 1.1 and 1.2 to give a positive number simultaneously for a particular application. If False Alarms and Misses are both positive, then Eq. 1.3 gives a positive value for False Alarm Exchange, which equals Miss Exchange. The concept of Exchange creates pairs between False Alarms and Misses, where the number of pairs is equal to the smaller of False Alarms and Misses. Equation 1.4 gives the Difference Quantity as the sum of False Alarm Quantity and Miss Quantity, which is also the absolute value of the difference between Misses and False Alarms. Difference Quantity indicates the absolute difference between the size of Presence in **X** and the size of Presence in **Y**. Equation 1.5 gives the Difference Exchange as two times the minimum of False Alarms and Misses, which is the sum of False Alarm Exchange and Miss Exchange. Difference Exchange indicates how the allocation of Presence in **X** differs from the allocation of Presence in **Y**. The Difference between **X** and **Y** is the sum of False Alarms and Misses, as Eq. 1.6 expresses.

Equation 1.6 shows how Difference is the sum of its two components: Difference Quantity and Difference Exchange.

Figure 1.1a shows the sizes of Quantity and Exchange for Misses and False Alarms. False Alarm Quantity is zero while Miss Quantity is one because the Presence in **Y** is one larger than the Presence in **X**. Both False Alarm Exchange and Miss Exchange are two because two is the smaller of False Alarms and Misses. Difference Quantity is one because one is the absolute difference between the Presence in **X** and the Presence in **Y**. Difference Exchange is four because there are two pairs of simultaneous Miss and False Alarm. Difference is five, which is the sum of its two components of Quantity and Exchange.

This paragraph describes a way to envision the concepts of Quantity and Exchange. Consider the ten observations for **X** and **Y** in Fig. 1.1. Suppose your job is to diagnose the ten observations in **Y** while the only thing you know about **Y** is that there are ten observations and each observation is either Presence or Absence. Your diagnosis is a combination of two types of decisions. One decision concerns how many Presences to diagnose. This first decision concerns Quantity. The second decision concerns how to allocate the diagnosed Presences among the observations. This second decision concerns allocation. In Fig. 1.1, the diagnosed quantity of Presence in **X** is 3, while the true quantity in **Y** is 4, thus the diagnosis is erroneous concerning quantity by 1 observation too few. Furthermore, the diagnosis in Fig. 1.1 is not optimal concerning allocation. An optimal allocation would maximize the number of Hits given the quantities in **X** and **Y**, meaning given the marginal sums in the square contingency table. The **X** variable diagnoses Presence in the three observations on the left at the top of Fig. 1.1. However, an optimal allocation of three diagnoses of Presence would have generated three Hits. It is possible to reallocate pairs of Presence and Absence in **X** to reduce differences with **Y**. Each pair of reallocation in **X** would convert one Miss into one Hit and one False Alarm into one Correct Rejection. There are two such pairs in Fig. 1.1. Reallocation in **X** could convert four observations from wrong to correct, thus Difference Exchange is four. After all such reallocations, there would still exist one erroneous observation, thus Difference Quantity is one. For another figure to explain the concepts of Quantity and Exchange, see Fig. 1 in Pontius Jr and Millones (2011).

A temptation can be to report proportion correct as the sum of Hits and Correct Rejections divided by the sum of all observations. Proportion correct is popular because it has a straightforward interpretation and can seem relevant initially. However, proportion correct might be misleading depending on the specific research question. For example, consider when we want to compare the ability of two sensors, **X** and **Z**, to detect true fires as **Y** describes. Sensor **X** produces the data in Fig. 1.1, meaning sensor **X** diagnoses fires for three of ten observations. The sensor **Z** is broken so sensor **Z** never diagnoses Presence and always diagnoses Absence. The Proportion Correct is 0.5 for **X** and 0.6 for **Z**, which would seem to suggest initially that the broken sensor **Z** is better than **X** for diagnosis. This illustrates why it is essential to select a metric that aligns with a particular research question. Some ratios are better suited than proportion correct to express diagnostic power. The literature contains numerous summary metrics that combine Misses, Hits, False

Alarms, and Correct Rejections into a single metric (Fielding and Bell 1997). The challenge is to select a metric that answers a relevant and specific question. Seventeen metrics are at https://en.wikipedia.org/wiki/Sensitivity_and_specificity. The concepts in this chapter are novel in the respect that I have not seen in the other literature a list of metrics that include the components of Quantity and Exchange. Meanwhile, the components of Quantity and Exchange are helpful to interpret most of the applications that I see.

This chapter's remaining equations give metrics in terms of intensities, which are ratios that answer common research questions. Figure 1.1b gives a way to envision results in terms of intensities. Equations 1.7, 1.8 and 1.9 give intensities of False Alarms, where the denominator in each equation is the size of Presence in **X**. False Alarm Quantity Intensity plus False Alarm Exchange Intensity equals False Alarm Intensity. Equations 1.10, 1.11 and 1.12 give intensities of Miss, where the denominator in each equation is the size of Presence in **Y**. Miss Quantity Intensity plus Miss Exchange Intensity equals Miss Intensity. Equations 1.13, 1.14 and 1.15 give intensities of differences, where the denominator in each equation is the size of the extent. Difference Quantity Intensity plus Difference Exchange Intensity equals Difference Intensity.

$$\text{False Alarm Quantity Intensity} = \frac{\text{MAXIMUM}\left(0, F - M\right)100\%}{H + F} \tag{1.7}$$

$$\text{False Alarm Exchange Intensity} = \frac{\text{MINIMUM}\left(M, F\right)100\%}{H + F} \tag{1.8}$$

$$\text{False Alarm Intensity} = \frac{F\,100\%}{H + F} \tag{1.9}$$

$$\text{Miss Quantity Intensity} = \frac{\text{MAXIMUM}\left(0, M - F\right)100\%}{M + H} \tag{1.10}$$

$$\text{Miss Exchange Intensity} = \frac{\text{MINIMUM}\left(M, F\right)100\%}{M + H} \tag{1.11}$$

$$\text{Miss Intensity} = \frac{M\,100\%}{M + H} \tag{1.12}$$

$$\text{Difference Quantity Intensity} = \frac{\left|M - F\right|100\%}{M + H + F + C} \tag{1.13}$$

$$\text{Difference Exchange Intensity} = \frac{2\,\text{MINIMUM}\left(M, F\right)100\%}{M + H + F + C} \tag{1.14}$$

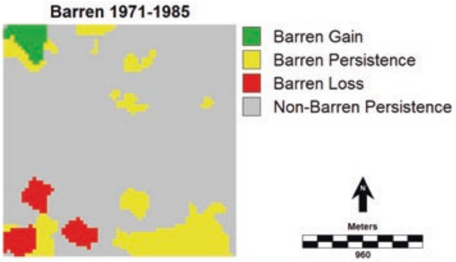
$$\text{Difference Intensity} = \frac{\left(M + F\right)100\%}{M + H + F + C} \tag{1.15}$$

False Alarm Intensity expresses False Alarms as a percentage of the Presence in **X**. Miss Intensity expresses Misses as a percentage of the Presence in **Y**. Difference Intensity expresses difference as a percentage of the extent. In the house fire example, Difference Intensity expresses the errors as a percentage of the number of houses. False Alarm Intensity expresses how frequently the detectors generate False Alarms as a percentage of houses in which the detectors diagnose fire. Miss Intensity expresses how frequently the detectors generate Misses as a percentage of houses in which true fires exist. It is helpful to compare the Difference Intensity to both False Alarm Intensity and Miss Intensity. Figure 1.1b shows that the False Alarm Intensity is greater than the Difference Intensity. This means that the detectors make errors more intensively in houses where the detectors diagnose fire than in all houses. Figure 1.1b shows that the Miss Intensity is greater than the Difference Intensity. This means that the detectors make errors more intensively in houses where fires exist than in all houses. Figure 1.1b shows that detectors generate errors more intensively when true fire is present than when true fire is absent, which is helpful information that proportion correct alone fails to reveal.

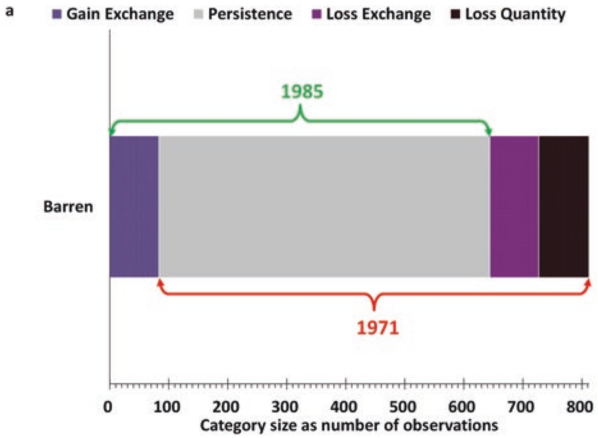
Figure 1.1b shows also that each type of difference is the sum of its two components: Quantity and Exchange. The Quantity component is smaller than the Exchange component for Misses, False Alarms, and Difference. The interpretation is that the detectors are more erroneous at diagnosing the allocation of fires among the houses than at diagnosing the quantity of fires in the extent.

Figure 1.2 gives an example to show how the concepts of this chapter analyze temporal change. The maps show Presence versus Absence of a land cover category called Barren. Each observation is a pixel that is 30 meters by 30 meters. Variable **X** is at the year 1971 while **Y** is at the year 1985. The maps show that most of Barren at 1971 persists from 1971 to 1985. Barren loses three patches in the southwest and gains one patch in the northwest. Figure 1.2a shows a Venn diagram with Barren at 1985 as the set on the left and Barren at 1971 as the set on the right. Persistence is the intersection of the two sets. Some pixels show Barren's Presence at 1971 and Absence at 1985, which is Barren's loss. Other pixels show Barren's Absence at 1971 and Presence at 1985, which is Barren's gain. Loss Quantity is positive and Gain Quantity is zero because Barren loses more than Barren gains. The Exchange components are positive because Barren loses in some pixels and gains in other pixels. Figure 1.2b shows the intensities. The extent bar indicates the percentage of the extent that experiences change is 6, which is the sum of 2 percentage points of the Quantity component and 4 percentage points of the Exchange component. The loss intensity indicates that Barren lost 23% of its start size. The gain intensity indicates that Barren's gain during 1971–1985 accounts for 13% of its end size.

Figure 1.1 shows a square contingency table that has nine entries, consisting of four central entries and five marginal sums. Figure 1.1 shows how the five marginal sums can derive mathematically from Hits, Misses, False Alarms, and Correct Rejections. It is tempting to think that the four central entries cause the marginal sums. However, causation in the table depends on the application. Furthermore, other combinations of four entries could generate the nine entries in the contingency table. Table 1.2 gives a sequence of four entries that are more helpful to consider



		1985		
		Barren	Non-Barren	Sum
1971	Barren	Persistence = 561	Loss = 168	729
	Non-Barren	Gain = 83	3284	3367
	Sum	644	3452	4096



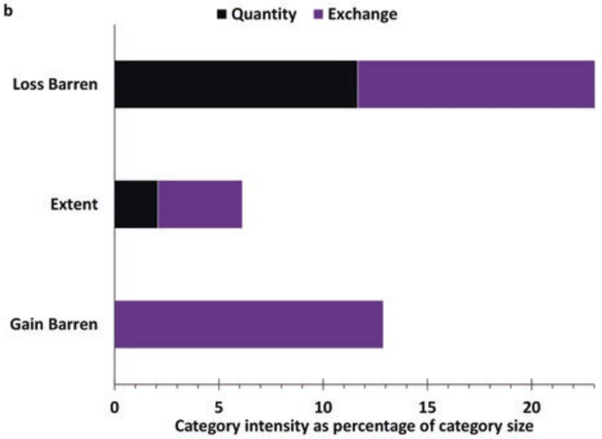


Fig. 1.2 Results for temporal change for Barren category

Table 1.2 Four entries that cause the other five entries in a table for applications to diagnosis

Entry	Mathematical expression using Table 1.1
Extent	Hits + Misses + False Alarms + Correct Rejections
Presence in **Y**	Hits + Misses = Abundance
Presence in **X**	Hits + False Alarms
Hits	Hits

when thinking about the construction of the contingency table. A common first step in scientific analysis is to determine the number of observations. The number of observations is the size of the extent, which is the first entry that Table 1.2 gives. The size of the extent is critical to interpret any of the results. A properly designed statistical analysis considers the extent size and then determines the sample size before collecting data, which allows computation of the sizes of Hits, False Alarms, Misses, and Correct Rejections. The second bit of critical information to understand diagnostic power is the size of Presence in truth, known as Abundance, which is the second entry in Table 1.2. The truth determines the size of Presence in **Y**. The size of Absence in **Y** is equal to the size of the extent minus the size of Presence in **Y**. Therefore, the first two entries in Table 1.2 dictate the marginal sums in the bottom row of the contingency table. The bottom row of marginal sums exists even when a diagnosis does not exist. Thus, the bottom row constrains the sizes of Hits, False Alarms, Misses, and Correct Rejections. The third entry in Table 1.2 is the size of Presence in **X**, e.g. the number of observations diagnosed as Presence. The method of diagnosis influences the size of Presence in **X**, which reveals whether the number of Presence diagnoses are fewer than or greater than the number of Presence observations in truth. The size of Absence in **X** equals the size of the extent minus the size of Presence in **X**. Thus, the top three measurements in Table 1.2 dictate all the marginal sums in the contingency table. The fourth entry in Table 1.2 is Hits, which allows a scientist to use the marginal sums to compute the sizes of False Alarms, Misses, and Correct Rejections. It is more helpful to think in terms of the sequence of the four entries in Table 1.2 than in Table 1.1 when analyzing diagnoses. Table 1.2 shows a sequence of influences that occur in diagnosis. The size of the extent constrains the size of Presence in **Y**, where both sizes exist independent of the diagnosis. The size of the extent constrains also the size of Presence in **X**, which in turn constrains the size of Hits. The size of Presence in **Y** constrains the size of Hits, while the size of Hits does not cause the size of Presence in **Y**. Neither the size of Presence in **X** nor the size of Presence in **Y** cause the size of the extent. These concepts play a central role in the next chapter.

The spreadsheet PontiusMatrix42.xlsx computes the equations for this chapter and generated the bar graphs in Figs. 1.1 and 1.2 (Pontius Jr 2020). The user enters the contingency table into the Excel file's Input sheet. Then the spreadsheet computes the numerical results and presents them in graphical form. PontiusMatrix42.xlsx is available for free from www.clarku.edu/~rpontius. The diffeR package in the software R reads raster maps to compute Quantity and Exchange (Pontius Jr and Santacruz 2015)

1.2 Discussion Questions

1. Which four entries cause the other five entries in a contingency table for applications to temporal change where **X** is the start time and **Y** is the end time?
2. What are the practical interpretations of Miss Intensity and False Alarm Intensity for applications that diagnose a disease?
3. What are some applications for which proportion correct is a misleading metric?
4. What are some applications where it is helpful to separate Difference into its two components of Quantity and Exchange?

References

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation, 24*, 38–49. https://doi.org/10.1017/S0376892997000088.

Pontius Jr, R. G. (2020). *PontiusMatrix42.xlsx*. http://www.clarku.edu/~rpontius

Pontius Jr, R. G., & Millones, M. (2011). Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing, 32*, 4407–4429. https://doi.org/10.1080/01431161.2011.552923.

Pontius Jr, R. G., & Santacruz, A. (2015). *diffeR: Metrics of difference for comparing pairs of maps*. https://cran.r-project.org/web/packages/diffeR

Chapter 2
Binary Variable Versus Rank Variable

Abstract This chapter gives methods to compare a variable that shows binary Presence or Absence versus a variable that indicates the ranked priority for Presence. The method considers several thresholds to convert the rank variable to a binary variable, which reduces each threshold to the analysis of the binary versus binary case of Chap. 1. The method presents the results graphically as the Total Operating Characteristic, which is more informative than the outdated Relative Operating Characteristic. Relevant software includes the TOC Curve Generator (Liu 2020) available via https://lazygis.github.io/projects/TOCCurveGenerator and R's TOC package available at https://cran.r-project.org/web/packages/TOC/index.html (Pontius Jr et al. 2015).

Keywords Relative Operating Characteristic · ROC · Total Operating Characteristic · TOC

2.1 Text

A binary variable gives for each observation exactly one of two possible states: Presence or Absence. A rank variable gives the sequence of observations in order of the priority to diagnose each observation as Presence. An observation's rank is a positive integer that denotes the observation's order in a sequence from first to last priority. This chapter gives methods to compare two variables when variable **Y** is a binary variable while variable **X** is a rank variable or any variable that can generate a rank variable. Consider using smoke density as a variable to diagnose fire's Presence or Absence for each observation. It would make sense to use smoke density to create a rank variable that orders a sequence for priority to diagnose fire with the highest smoke density first and the lowest smoke density last. Any interval or ratio variable can generate a rank variable. If a categorical variable contains an ordering, such as High, Medium, and Low, then the categorical variable can generate a rank variable where the observations in each category have a tied rank.

Figure 2.1 illustrates the concepts with example data. The **Y** variable is a binary variable, thus has two states: Presence or Absence. For the applications in this chapter, we define Presence as the state that exists for fewer than half of the **Y**

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1_2

X	90	65	50	45	40	30	30	30	10	10
Y	P	A	A	P	P	P	A	A	A	A

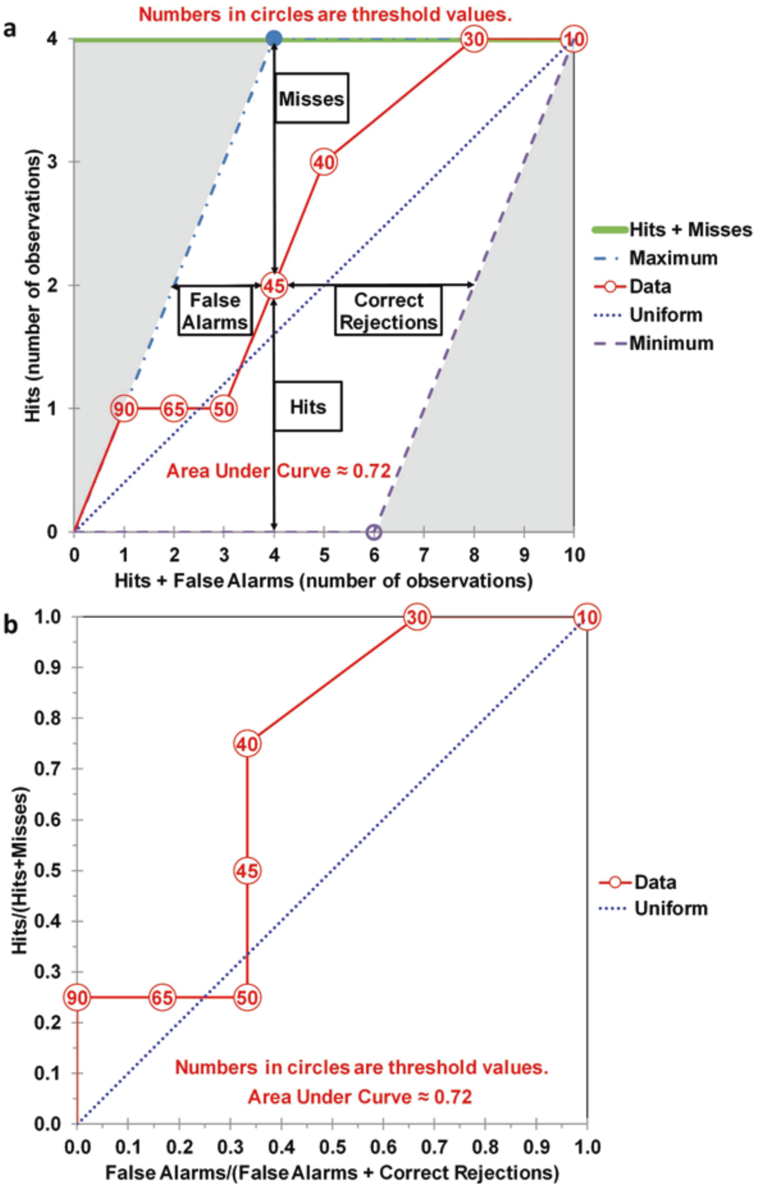


Fig. 2.1 Example where **X** generates ranks while **Y** distinguishes Presence (P) from Absence (A)

observations. If exactly half of the **Y** observations are one state, then it does not matter mathematically which state is Presence. The variable **Y** in Fig. 2.1 is identical to **Y** in Fig. 1.1, where the observations consist of four Presence and six Absence. The **X** variable in Fig. 2.1 generates a ranked priority in which we would diagnose each observation as Presence with earlier priority for larger **X** values. Figure 2.1 shows larger numbers at the left, thus the sequence of priority to diagnose observations as Presence is first at the left and last at the right. The rank dictates the sequence of priority for allocation of Presence, while the rank does not give information concerning the quantity of Presence.

We consider several possible thresholds to diagnose each observation as either Presence or Absence. When larger **X** values have earlier priority, if an observation's **X** value is greater than or equal to a threshold, then we diagnose the observation as Presence, otherwise as Absence. If we were to select a threshold of 50 in the example, then the diagnosis would be identical to **X** in Fig. 1.1. Each possible threshold converts the rank variable into a binary variable to compare with **Y**. Then the concepts of Chap. 1 apply to compare two binary variables because each threshold generates a contingency table that gives Hits, False Alarms, Misses, and Correct Rejections.

The Total Operating Characteristic (TOC) is an informative method to consider various thresholds when comparing the ranked values of the **X** variable to the binary **Y** variable (Pontius Jr and Si 2014). Two axes define the TOC space. Both axes show sizes in terms of the number of observations. The horizontal axis shows the size of the diagnosed Presence, which depends on the threshold. The size of the diagnosed Presence is the sum of Hits and False Alarms. The horizontal axis ranges from zero to the extent's size, meaning the number of observations. The vertical axis is the size of Hits, which depends on the threshold. An observation must have Presence in **Y** for a threshold's diagnosis to generate a Hit. Therefore, the vertical axis ranges from zero to the size of Presence in **Y**, known as Abundance, which is the sum of Hits and Misses. The size of Presence in **Y** is by design less than or equal to half of the size of the extent, therefore the maximum value on the vertical axis is less than or equal to half the maximum value on the horizontal axis. The TOC space is square visually, meaning the length of the vertical axis is identical to the length of the horizontal axis.

Let us walk through the example data to understand how each threshold generates a point on the TOC curve in Fig. 2.1a. The maximum **X** value indicates the first priority to diagnose Presence for a situation where larger **X** values have earlier priority. If the scientist selects a threshold greater than the maximum value in **X**, then the diagnosis is Absence for all observations. This initial threshold causes zero Hits and zero False Alarms, thus generates a point at the origin (0,0) of the TOC space, which is always the beginning of any TOC curve. Then the scientist considers incremental modifications to the threshold. The next threshold is the maximum **X** value, which is 90. The threshold at 90 diagnoses the first **X** observation as Presence, which generates in one Hit and zero False Alarms. Thus, the TOC curve climbs from the origin up the one-to-one line, where Hits equals the sum of Hits and False Alarms. Then the scientist lowers the threshold to the next smaller **X** value, which is 65. The threshold at 65 diagnoses the first two **X** observations as Presence, which generates one Hit and one False Alarm. Thus, the TOC curve extends one additional observation horizontally to the right. The subsequent threshold of 50 diagnoses the first three **X** observations as Presence, which generates one Hit and two False Alarms, thus the TOC curve extends horizontally an additional observation farther to the right. The next thresholds are 45 and 40, which generate two additional Hits and zero additional False Alarms, thus the TOC curve climbs parallel to the one-to-one line, which is the left boundary of the TOC parallelogram. The next three observations have **X** values tied at 30, where **Y** shows one of those observations as Presence and two as Absence. If the threshold is 30, then diagnosed Presence is eight, which consists of four Hits and four False Alarms. The last threshold is the minimum **X** value, which is 10. The last threshold diagnoses all observations as Presence, thus all observations are either Hits or False Alarms. At the last threshold, Hits equals the size of Presence in **Y**, at which point both Misses and Correct Rejections are zero. The last threshold always generates a point at the upper right corner of the TOC space.

The TOC space has maximum and minimum bounds in which the TOC curve resides. The size of Presence in **Y** dictates the bounds. The maximum bound portrays a perfectly correct TOC curve, which means all of the Presence observations in **Y** correspond to **X** values that have ranks earlier in the sequence of thresholds. Many **X** variables could portray a perfectly correct TOC curve. The maximum bound begins at the origin and then climbs along the one-to-one line to the point where the diagnosed quantity matches the quantity of Presence in **Y**. The filled circle in Fig. 2.1a denotes that point, which is the only point where all observations are correct. Then the increase in diagnosed quantity generates False Alarms, which causes the maximum bound to progress horizontally to the final point at the upper right corner where all observations diagnose Presence. The minimum bound portrays a perfectly erroneous TOC curve, which means all of the Absence observations in **Y** correspond to **X** values that have ranks earlier in the sequence of thresholds. Many **X** variables could portray a perfectly erroneous TOC curve. The minimum bound begins at the origin and then follows the horizontal axis to the point where the diagnosed quantity matches the quantity of Absence in **Y**, which is the size of the extent minus the size of Presence in **Y**. The unfilled circle in Fig. 2.1a denotes that point, which is the only point where all observations are erroneous. Then the increase in diagnosed quantity generates Hits to the final point at the upper right corner where all observations diagnose Presence. The maximum and minimum bounds form a parallelogram. Any TOC curve cannot have points above the maximum bound or below the minimum bound, which is why those spaces are gray. The maximum and minimum bounds form a helpful frame of reference. Another helpful reference is the straight diagonal line between the origin and the upper right corner of the TOC space, which Fig. 2.1 denotes as the uniform line. If all the **X** values were the same number, then all observations would have a tied rank, in which case the TOC curve would be that diagonal uniform line. If the **X** values were random numbers, then the mathematically expected TOC curve would be the diagonal uniform line.

Each threshold generates a point on the TOC curve; and each point on the TOC curve shows a threshold's Hits, False Alarms, Misses, and Correct Rejections. Figure 2.1a highlights the point at the threshold where the diagnosed quantity matches the quantity of Presence in **Y**. Hits is the vertical distance between a point on the TOC curve and the horizontal axis. False Alarms is the horizontal distance between a point and the left maximum bound. Misses is the vertical distance between a point and the horizontal line that denotes Hits plus Misses. Correct Rejections is the horizontal distance between a point and the right minimum bound.

Scientists should show the TOC curve and interpret its shape in various regions of the TOC parallelogram in the context of a particular research question. The shape of the curve shows where the Presence of **Y** is concentrated more or less intensely than uniform. Presence in **Y** is more intensive than uniform between two thresholds when the slope between the two thresholds on the TOC curve is steeper than the uniform line. Presence in **Y** is less intensive than uniform between two thresholds when the slope between the two thresholds on the TOC curve is flatter than the uniform line. Some regions of the TOC space might be more important than other regions depending on the research question. Important regions contain the thresholds that a decision-maker would consider to make a practical decision. The TOC curve between the origin and the size of Presence in **Y** shows the earlier thresholds, which might be the thresholds where realistic options for important decisions reside. In that case, the part of the TOC curve near the origin or near the size of Presence in **Y** would be more important than the part of the TOC curve near the upper right corner. The upper right corner of the TOC space might represent thresholds that are not interesting for practical decisions. Scientists should interpret the curve's steepness near the important thresholds and the curve's overall shape.

The TOC shows for each threshold the total information necessary to fill the threshold's contingency table in the format of Fig. 1.1. Specifically, the TOC shows the sizes of Hits, False Alarms, Misses, and Correct Rejections. The TOC shows also the entries in Table 1.2. The size of the extent is the maximum value on the horizontal axis, the size of Presence in **Y** is the maximum value on the vertical axis, the size of Presence in **X** is the horizontal coordinate of a threshold's point on the TOC curve, and Hits is the vertical coordinate of a threshold's point on the TOC curve. The TOC is an effective way to describe the contingency tables for several thresholds simultaneously in a single graph. A less informative method is the popular Relative Operating Characteristic (ROC). Figure 2.1 compares the TOC in part a to the ROC in part b for the same data. The algorithm to generate a ROC curve follows the same logic to generate the TOC curve, meaning the ROC considers a ranked sequence of thresholds for the **X** variable to diagnose Presence or Absence (Fawcett 2006; Swets 1988). The ROC curve plots the results on axes that are less informative than the axes of the TOC. The axes of the TOC show sizes, whereas the axes of ROC show relative intensities. ROC's horizontal axis is a unitless ratio of False Alarms to the sum of False Alarms and Correct Rejections. ROC's vertical axis is a unitless ratio of Hits to the sum of Hits and Misses. There is a one-to-one correspondence between the points on a TOC curve and the points on the ROC curve. However, the ROC curve gives insufficient information to reveal the size of the entries in the contingency table at each threshold, because each point on the ROC curve gives two unitless ratios that range from zero to one. The ROC curve fails to show crucial information that the TOC shows clearly, such as the size of the extent, the size of the Presence in **Y**, the threshold that diagnoses the correct size of Presence in **Y**, and each threshold's size of Hits, False Alarms, Misses, and Correct Rejections. The TOC contains sufficient information to generate the ROC; however, the ROC lacks sufficient information to generate the TOC. The ROC fails to convey two crucial bits of information: the size of the extent and the size of the Presence in **Y**. Those two bits of information are essential for insightful interpretation, while they are independent of any diagnosis. Those two bits of information would allow us to transform a ROC into its corresponding TOC. The TOC shows strictly more information than the ROC, which is why scientists should use the TOC rather than the ROC. ROC curves have been popular in a variety of professions where TOC curves would have been more informative. If authors would publish TOC curves, then readers could use TOC curves to make informed decisions. Let us consider two examples. One from the fire alarm example of Chap. 1 and another from medicine.

Consider a device that senses smoke density to diagnose the Presence or Absence of fire in houses. Denser smoke is greater evidence of fire's Presence. The engineer who creates the alarm must decide how to select the threshold of smoke density that triggers the device to diagnose fire. Each possible selection of a threshold generates numbers of Hits, False Alarms, Misses, and Correct Rejections. The engineer wants to select the optimal threshold, which requires a definition of optimal. The definition of optimal should consider the costs of Misses relative to the costs of False Alarms. A False Alarm occurs when the device signals fire when fire is absent, in which case the False Alarm annoys the house's residents. On the other hand, a Miss occurs when fire is present but the device fails to signal the fire, in which case the house's residents could die. A Hit is likely to save the lives of the residents. A Correct Rejection would be equivalent to the residents living without the device while fire is absent. In this example, the cost of a Miss is greater than the cost of a False Alarm, in which case an optimal threshold would allow more False Alarms than Misses. Although in practice, if the False Alarms are too frequent, then the residents might learn to ignore the device, which could leave the residents as vulnerable as having no device. Sensitivity is the ratio of Hits to the sum of Hits and Misses. The engineer wants to assure a high Sensitivity, implying a high probability of the alarm sounding, given that a real fire exists. Therefore, the engineer designs the device to be very sensitive, which as a side effect causes a substantial number of annoying False Alarms.

As another example, consider a doctor who uses a screening test that measures the concentration of a chemical in a patient's blood to diagnose a disease's Presence or Absence. The doctor considers higher concentrations of the chemical as stronger evidence of the disease's Presence. The doctor wants to select the optimal threshold of concentration to diagnose Presence. A threshold at a lower concentration would lead to more diagnoses of the disease's Presence, which could reduce Misses but increase False Alarms. A False Alarm occurs when the doctor diagnoses the disease's Presence when the patient is healthy, in which case the patient might undergo unnecessary treatment, which could have painful side effects. A Miss occurs when the doctor diagnoses the disease's Absence when the patient has the disease, in which case the patient could suffer from the untreated disease. If the disease is life-threatening, while the treatment is not painful, then the optimal threshold would allow more False Alarms than Misses. If the disease is a mere inconvenience, while the treatment is painful, then the optimal threshold would allow more Misses than False Alarms. The doctor might wonder how useful the concentration of a chemical in a patient's blood is to diagnose the disease and whether other variables have stronger diagnostic power. The doctor needs a method to compare among various **X** variables. If the doctor has numerous **X** variables, then it would be helpful to have a summary metric to sort the various **X** variables in terms of overall diagnostic power. The Area Under the Curve (AUC) is an appropriate metric for the initial sort.

The AUC, pronounced as awk, is a metric that synthesizes across thresholds the ability of the ranked values of **X** to diagnose the allocation of Presence in **Y**. A larger AUC indicates a greater ability of the ranked values of **X** to diagnose the allocation of Presence in **Y**. AUC is a metric that ranges from zero to one, where zero means the diagnoses from **X** are perfectly erroneous and one means the diagnoses from **X** are perfectly correct. The TOC shows the AUC as a ratio, where the numerator is the area under the TOC curve that is in the bounding parallelogram, and the denominator is the area of the bounding parallelogram. The AUC of the TOC's maximum bound is one. The AUC of the TOC's minimum bound is zero. The AUC of the uniform line is 0.5. If the **X** values were random numbers, then the expected AUC would be 0.5. Thus, an AUC of 0.75 indicates the diagnostic ability of **X** to allocate the Presence of **Y** is halfway between random and perfect. The AUC equals 0.72 in Fig. 2.1, which means that the ability of **X** to diagnose the allocation of Presence in **Y** is less than halfway between random and perfect. The end of this chapter shows mathematically how to compute the AUC.

The AUC measures the strength of a monotonic association between the Presence in **Y** and the ranks that derive from **X**. AUC values greater than 0.5 indicate a positive association between Presence in **Y** and ranks earlier in the sequence for **X**. AUC values less than 0.5 indicate a negative association between Presence in **Y** and ranks earlier in the sequence for **X**. An AUC of 0.5 indicates lack of monotonic relationship between the Presence in **Y** and the ranks for **X**. If the TOC curve is below the uniform line at some thresholds and above the uniform line at other thresholds, then there exists a non-monotonic association between the Presence in **Y** and the ranks of **X**, in which case the AUC might be 0.5. This illustrates the danger of using a single number such as the AUC to judge an overall relationship. There might be an important non-monotonic relationship between the Presence in **Y** and the ranks of **X**, which the shape of the TOC curve would show, but which the AUC would not necessarily convey.

Some authors are tempted to apply universal rules to designate various ranges of AUC as poor, acceptable, good, excellent, et cetera. However, any universal rule does not address any particular research question or specific application, precisely because the rule is universal. Universal rules serve more to address scientists' psychological desires rather than any scientific purposes. Universal rules are dangerous when the rules cause scientists or their audience to stop thinking after obtaining an AUC that the scientist claims is acceptable. Furthermore, universal rules discourage scientists from deciding the level of acceptability for each particular application. It can be complicated to define acceptability for any particular application because there are frequently many factors to consider. Any rule to designate a metric's value as acceptable must depend on a definition of acceptable for a particular purpose. For example, if the purpose is to determine whether an **X** variable is acceptable to diagnose a disease, then the determination might mean the difference between life and death. If the purpose is to determine whether the softness of a bicycle tire is acceptable to diagnose low air pressure, then the stakes of the diagnosis are lower than the stakes of the diagnosis of a deadly disease. The disease and the tire differ concerning the expense of the diagnostic test and in the importance of the resulting diagnosis. It makes no sense to use one universal value of AUC to define acceptability for both cases. Moreover, it is not necessary to define acceptability to obtain valuable insight.

It is frequently helpful to compute a baseline AUC to help to interpret the AUC that derives from a particular **X** variable. A baseline AUC must relate to the specific application. For example, if the application relates to a newly-proposed diagnostic variable **X**, then the baseline AUC should relate to the previously-used diagnostic variable. Some scientists are tempted to use the uniform line and its AUC of 0.5 as the baseline for comparison with a newly-proposed variable. The uniform line portrays a random ranking of observations, which is irrelevant for most of the practical applications that I have seen. For example, doctors do not diagnose diseases randomly, even when diagnostic variables are unavailable. An AUC greater than 0.5 could exist for a simple diagnostic variable such as age, for diseases that tend to affect older patients. For example, a doctor would rank older patients before younger patients for the diagnosis of prostate cancer (Swets et al. 2000). Thus, a relevant research question is whether a newly-proposed variable could generate an AUC greater than the AUC that derives from the single variable of age. Another example is a geographer who wants to explain the allocation of deforestation during a time interval. Experienced geographers spend substantial effort considering several variables, while a naïve geographer could assume that deforestation is more intensive nearer the deforestation that had occurred during a preceding time interval (Pontius Jr 2018). The naïve explanation is likely to have an AUC substantially larger than 0.5 because humans do not deforest randomly (Pontius Jr and Batchu 2003; Pontius Jr and Si 2014). Then the question would be whether additional variables offer stronger explanatory power than the naïve explanation. If scientists interpret the shape of the TOC and use the AUC intelligently with respect to a relevant baseline, then the TOC and its AUC can be helpful to see the implications of using various possible **X** variables.

Some authors have criticized the use of the AUC for a variety of legitimate reasons (Cook 2007; Golicher et al. 2012; Lobo et al. 2008; Peterson et al. 2008). However, any metric gives only a single bit of information. The AUC measures exactly what it promises, which is a single metric that summarizes the curve across its thresholds in terms of the area under the curve. AUC does not promise to

Table 2.1 Notation to compute the Area Under Curve for TOC and ROC

Notation	Meaning
C_t	Size of Correct Rejections at threshold t
F_t	Size of False Alarms at threshold t
H_t	Size of Hits at threshold t
M_t	Size of Misses at threshold t
t	Index for threshold, where $t = 0, 1, 2, \ldots, T$
T	Index for the last threshold at the upper right corner of the TOC space

measure the details of the shape of the TOC or ROC curves. If scientists report only the AUC when the AUC is irrelevant or when other metrics are also relevant, then that is the fault of the scientists, not of the AUC.

Table 2.1 gives the mathematical notation to compute the AUC. The subsequent equations prove that the AUC of the TOC equals the AUC of the ROC. This illustrates again how the TOC contains strictly more information than the ROC, while the TOC maintains the properties that have made ROC so popular.

Free software exists to compute the TOC. The TOC package in the software R can read raster maps to compute and plot the TOC (Pontius Jr et al. 2015). The TOC Curve Generator has a variety of helpful features that do not exist in other software packages (Liu 2020).

$$\text{Area Under Curve for Total Operating Characteristic} =$$

$$\frac{\{\text{area under TOC curve}\} - (\text{area of lower right triangle outside TOC parallelogram})}{\text{area of TOC parallelogram}} =$$

$$\frac{\left\{\sum_{t=1}^{T} \text{area of TOC trapezoid}_t\right\} - (\text{area of right triangle outside TOC parallelogram})}{\text{area of TOC parallelogram}} =$$

$$\frac{\left\{\sum_{t=1}^{T}\left[(\text{width of TOC trapezoid}_t)(\text{average of legs of TOC trapezoid}_t)\right]\right\} - \left(H_T^2/2\right)}{(\text{base TOC parallelogram})(\text{perpindicular height of TOC parallelogram})} =$$

$$\frac{\left\{\sum_{t=1}^{T}\left[(H_t + F_t) - (H_{t-1} + F_{t-1})\right]\left[(H_t + H_{t-1})/2\right]\right\} - \left(H_T^2/2\right)}{(F_T)(H_T)} =$$

$$\frac{\left\{\sum_{t=1}^{T}\left[(H_t - H_{t-1}) + (F_t - F_{t-1})\right](H_t + H_{t-1})\right\} - H_T^2}{2F_T H_T} =$$

$$\frac{\left\{\sum_{t=1}^{T}\left[\left(H_t-H_{t-1}\right)\left(H_t+H_{t-1}\right)+\left(F_t-F_{t-1}\right)\left(H_t+H_{t-1}\right)\right]\right\}-H_T^2}{2F_TH_T}=$$

$$\frac{\left\{\sum_{t=1}^{T}\left[\left(H_t^2-H_{t-1}^2\right)+\left(F_t-F_{t-1}\right)\left(H_t+H_{t-1}\right)\right]\right\}-H_T^2}{2F_TH_T}=$$

$$\frac{\left\{\sum_{t=1}^{T}\left[H_t^2-H_{t-1}^2\right]\right\}+\left\{\sum_{t=1}^{T}\left[\left(F_t-F_{t-1}\right)\left(H_t+H_{t-1}\right)\right]\right\}-H_T^2}{2F_TH_T}=$$

$$\frac{H_T^2+\left\{\sum_{t=1}^{T}\left[\left(F_t-F_{t-1}\right)\left(H_t+H_{t-1}\right)\right]\right\}-H_T^2}{2F_TH_T}=$$

$$\frac{\sum_{t=1}^{T}\left[\left(F_t-F_{t-1}\right)\left(H_t+H_{t-1}\right)\right]}{2F_TH_T}=$$

$$\sum_{t=1}^{T}\left[\left(\frac{F_t-F_{t-1}}{F_T}\right)\left(\frac{H_t+H_{t-1}}{2H_T}\right)\right]=$$

$$\sum_{t=1}^{T}\left[\left(\frac{F_t}{F_T}-\frac{F_{t-1}}{F_T}\right)\left(\frac{H_t}{2H_T}+\frac{H_{t-1}}{2H_T}\right)\right]=$$

$$\sum_{t=1}^{T}\left[\frac{F_t}{F_t+C_t}-\frac{F_{t-1}}{F_{t-1}+C_{t-1}}\right]\left[\left(\frac{H_t}{H_t+M_t}+\frac{H_{t-1}}{H_{t-1}+M_{t-1}}\right)/2\right]=$$

$$\sum_{t=1}^{T}\left[\text{width of ROC trapezoid}_t\right]\left[\text{average of legs of ROC trapezoid}_t\right]=$$

Area Under Curve for Relative Operating Characteristic

2.2 Discussion Questions

1. Which entries in a threshold's contingency table are independent of the diagnosis and what features of the TOC do they dictate?
2. What is the importance of the coordinates of the point at the upper left of the TOC parallelogram?
3. What is the importance of the first point on the TOC curve that touches the Maximum boundary?
4. What is the interpretation of the slope of the segment between two points on the TOC curve?

5. What characteristics are important when interpreting the shape of a TOC curve?
6. How many TOC curves could have the same AUC?
7. If a scientist wants to select an optimal threshold, then what factors should the scientist consider?
8. Under what conditions should you consider a particular value of AUC as acceptable?
9. What bits of information would you need to transform a ROC curve into a TOC curve?
10. What information does the TOC show that the ROC does not show?
11. Some authors plot TOC or ROC curves using software that smooths the curves and that lack threshold labels. How do such practices hinder interpretation?
12. If authors report the AUC values but do not show the corresponding TOC curves, then how might such reporting cause misinterpretation?
13. If Abundance were nearly equal to the extent's size, then how would the TOC parallelogram appear?

References

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation, 115*, 928–935. https://doi.org/10.1161/CIRCULATIONAHA.106.672402.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 14*. https://doi.org/10.1016/j.patrec.2005.10.010.

Golicher, D., Ford, A., Cayuela, L., & Newton, A. (2012). Pseudo-absences, pseudo-models and pseudo-niches: Pitfalls of model selection based on the area under the curve. *International Journal of Geographical Information Science, 26*, 2049–2063. https://doi.org/10.1080/13658816.2012.719626.

Liu, Z. (2020). *TOC Curve Generator*. https://lazygis.github.io/projects/TOCCurveGenerator

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography, 17*, 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x.

Peterson, A. T., Papes, M., & Soberon, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling, 213*, 63–72. https://doi.org/10.1016/j.ecolmodel.2007.11.008.

Pontius Jr, R. G. (2018). Criteria to confirm models that simulate deforestation and carbon disturbance. *Land, 7*, 14. https://doi.org/10.3390/land7030105.

Pontius Jr, R. G., & Batchu, K. (2003). Using the relative operating characteristic to quantify certainty in prediction of location of land cover change in India. *Transactions in GIS, 7*, 467–484. https://doi.org/10.1111/1467-9671.00159.

Pontius Jr, R. G., & Si, K. (2014). The total operating characteristic to measure diagnostic ability for multiple thresholds. *International Journal of Geographical Information Science, 28*, 570–583. https://doi.org/10.1080/13658816.2013.862623.

Pontius Jr, R. G., Santacruz, A., Tayyebi, A., & Parmentier, B. (2015). *TOC: Total operating characteristic curve and ROC curve*. https://cran.r-project.org/web/packages/TOC

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285–1293. https://doi.org/10.1126/science.3287615.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better DECISIONS through SCIENCE. *Scientific American, 283*, 82–87. https://www.scientificamerican.com/article/better-decisions-through-science/.

Chapter 3
Application of the Total Operating Characteristic

Abstract This chapter uses the Total Operating Characteristic (TOC) and empirical maps to describe the gain of the Built category during each of two time intervals in relation to two independent variables: initial land cover and distance from initial Built. The TOC curve shows that Built's gain during the first time interval is more intensive on the initial Barren category and nearer the initial Built. Area Under the Curve values indicate the relationships are weaker during the second time interval, when Built's gain is more intensive on the initial Barren category and farther from the initial Built.

Keywords Land change · Total Operating Characteristic · TOC

3.1 Text

The previous chapter explained how the Total Operating Characteristic (TOC) reveals the relationship between a binary variable and a rank variable (Pontius Jr and Si 2014). There are as many applications to TOC as there are for the popular Relative Operating Characteristic (ROC), meaning numerous applications across a variety of sciences (Swets et al. 2000). However, TOC gives more information than ROC in a plot that occupies the same amount of space. TOC applies to remote sensing where an investigator wants to know the degree to which information in an image distinguishes Presence from Absence of a category on the ground (Dustin and Jacobson 2015). Investigators have applied TOC to other applications, such as to test sensors that detect when elderly people fall (Alves et al. 2019). TOC is appropriate for comparing various algorithms that predict changes (Cushman et al. 2017; Shafizadeh-Moghadam et al. 2017a, b). The TOC curve show sizes such as numbers of observations, thus the TOC is a more effective way to communicate than summary metrics of ratios (Kamusoko and Gamba 2015). The TOC curve allows readers to interpret results clearly and to identify flaws in validation methodology, such as neglecting to distinguish between persistence and change when the purpose is to predict change (Chakraborti et al. 2018; Naghibi and Delavar 2016; Pontius Jr and Parmentier 2014). This chapter illustrates the use of TOC to describe relationships in land change science.

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic
Information Science, https://doi.org/10.1007/978-3-030-70765-1_3

This chapter uses the TOC to characterize the gain of a land cover category, specifically the Built category. The maps derive from the Bureau of Geographic Information for the State of Massachusetts in the USA, known as MassGIS (https://www.mass.gov/orgs/massgis-bureau-of-geographic-information). MassGIS supplies maps of 21 categories of land cover at the years 1971, 1985, and 1999. I aggregated those 21 categories into a smaller number of categories according to the Anderson classification system. The spatial extent that this chapter examines has four land cover categories: Built, Barren, Forest and Water. This chapter analyzes the gain of the Built category during two time intervals: 1971–1985 and 1985–1999. During each time interval, the TOC reveals the relationship between Built's gain and two variables. The first variable is the distance from Built at the start of the time interval. The second variable is the land cover category at the start of the time interval. The digital maps are in raster form, meaning each observation is a pixel. Each pixel is a square that is 30 m on a side. The spatial extent consists of 64 rows and 64 columns of pixels, thus each map has 4096 observations that cover 368.64 ha. The TOC has the ability to express the results in terms of the number of observations or the area for this application. The remainder of this chapter expresses the results in terms of hectares.

Figure 3.1 examines the time interval 1971–1985. The map in the upper left shows: Built at 1971, Built's gain during 1971–1985, and Non-Built Persistence during 1971–1985. Built at 1971 is not part of the extent that the TOC analyzes because land must be initially Non-Built in order possibly to experience gain of Built during the time interval. Presence for the binary variable is Built's gain during 1971–1985. Absence for the binary variable is Non-Built Persistence during 1971–1985. The sizes of the categories in the map in the upper left of Fig. 3.1 dictate the parallelogram that bounds the TOC. Gain of Built appears on 41 hectares of the 314 hectares of Non-Built at 1971. Thus, TOC has an upper bound of 41 hectares on the vertical axis and a right bound of 314 on the horizontal axis. The corners of the parallelogram that bounds the TOC are (0,0), (41,41), (314,41), and (273,0). The TOC curve for any independent variable must reside in or on those bounds. The uniform line portrays a hypothetical situation where the Built's gain has uniform intensity throughout the Non-Built at 1971. The uniform line gives a baseline to interpret the TOC curve for any independent variable.

The left side of Fig. 3.1 shows a map of the distance from Built at 1971, where black indicates Built at 1971 and lighter shades indicate distances farther from Built at 1971. The TOC shows the relationship between Built's Gain during 1971–1985 and distance from Built at 1971. A hypothesis is that Built's gain occurs more intensively nearer to Built at 1971. Thus, the sequence of the ranks of the distance variable prioritizes smaller distances first. The TOC curve measures whether Built's gain occurs more intensively on darker shades in the distance map. The TOC considers several thresholds to reclassify each observation of the distance variable as Presence or Absence of Built's gain. Each threshold generates a point on the TOC curve. The threshold's label at each point indicates the threshold's distance in meters from Built at 1971. The TOC curve begins at the origin of the TOC space. The sequence of thresholds progress in steps of 30 m, due in part to the spatial resolution of the data; the observations are pixels that have 30 m per side. The threshold at 30 m is the threshold where the quantity of the union of Hits and False Alarms is most similar to the quantity of the union of Hits and Misses, which is 41 ha. The map in the lower left of Fig. 3.1 shows the Hits, False Alarms, Misses, and Correct Rejections for the threshold at 30 m. The Hits and False Alarms are within one pixel of the Built at 1971, meaning within 30 m. The map in the lower left shows spatial arrangement concerning the distance between Misses and False Alarms, which the TOC fails to show.

Interpretation of the TOC curve offers a plethora of information. Every threshold of the TOC separates the extent into two parts: the part nearer to the Built at 1971 and the part farther from the Built at 1971. Every such threshold generates a point on the TOC curve that is above the uniform line, which indicates for all thresholds that Built's gain is more intensive in the part of the extent that is nearer to the Built at 1971. Each pair of consecutive thresholds form a bin that captures an incremental increase on the horizontal axis, meaning an incremental increase in the sum of Hits and False Alarms. Thus, each pair of thresholds forms a straight segment of the TOC curve, where the slope of the segment is a ratio of the increase in Hits to the increase in the sum of Hits and False Alarms in the bin. The slope indicates the intensity with which Hits occupy the bin. If a segment's slope is steeper than the uniform line's slope, then the intensity of Built's gain in the segment's bin is greater than the intensity of Built's gain in the Non-Built at 1971. If a segment's slope is flatter than the uniform line's slope, then the intensity of Built's gain in the segment's bin is less than the intensity of Built's gain in the Non-Built at 1971. The steepest segment of the TOC curve is between the thresholds at 30 and 60 m, which means that Built's gain is most intensive in the bin that is greater than 30 m and simultaneously less than or equal to 60 m. The segments' slopes are steeper than uniform in the segments from the origin of the TOC space up to the threshold at 210 m, which indicates that Built's gain is more intensive than uniform in bins nearest to Built at 1971. The segments' slopes are flatter than uniform in the segments beyond the threshold at 210 m, which indicates that Built's gain is less intensive than uniform in bins farthest from Built at 1971. The threshold at 540 meters is where the TOC curve meets the upper bound, which indicates that all of Built's gain exists within 540 meters of the Built at 1971. The Area Under the Curve (AUC) is 0.60, which summarizes the overall direction and strength of the relationship between Built's gain and ranked distance from Built at 1971. AUC is greater than 0.5, which indicates that Built's gain is overall more intensive at places nearer to Built at 1971. The AUC of 0.60 is closer to 0.5 than to 1, which indicates the strength of the relationship between Built's gain and ranked distance from Built at 1971 is closer to uniform than to perfect.

The right side of Fig. 3.1 analyzes the variable that derives from land cover at 1971. Land cover has four categories: Barren, Forest, Water and Built. Built is eliminated from the TOC analysis, because an observation must be Non-Built at 1971 in order to have the possibility to experience the gain of Built during 1971–1985. The TOC requires a ranking of the remaining three categories. I ranked the categories as Barren first, then Forest second, and Water last based on the intensity with which the

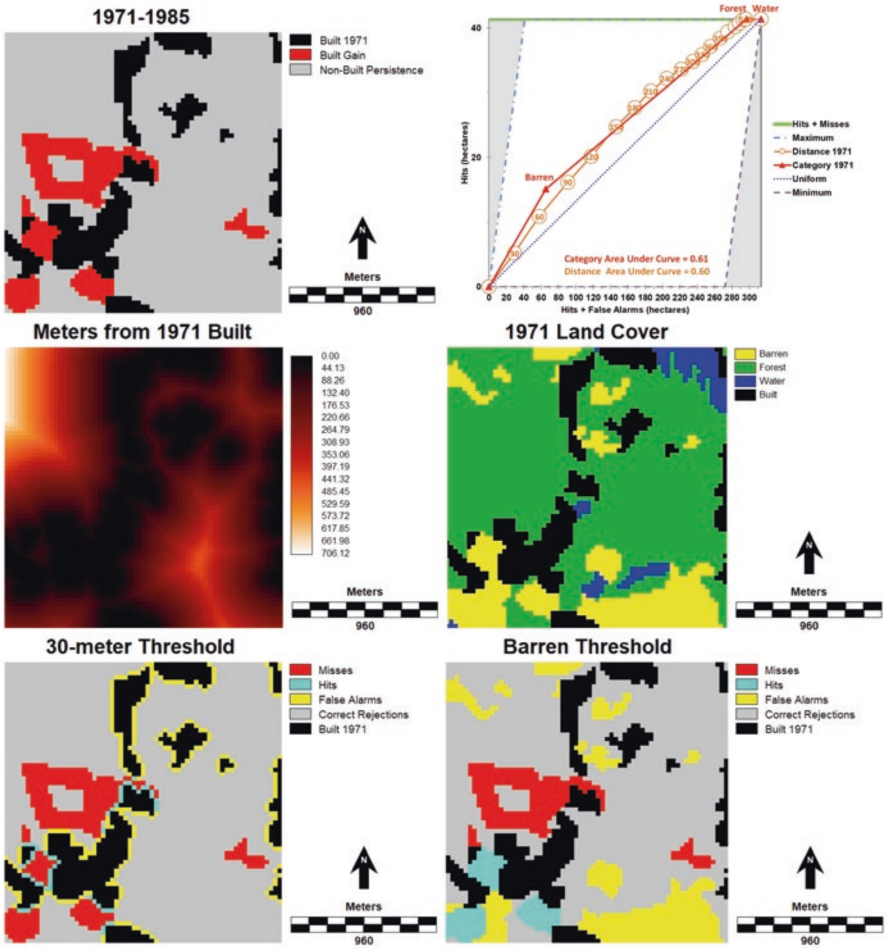


Fig. 3.1 Application of Total Operating Characteristic to describe Built's gain during 1971–1985

three categories experience the gain of Built. Each category's intensity is a ratio where the numerator is the size of Built's gain in the category and the denominator is the size of the category. Barren has the greatest intensity, as 23% of Barren at 1971 transitions to Built during 1971–1985. Forest has the second greatest intensity, as 11% of Forest at 1971 transitions to Built during 1971–1985. Water has the smallest intensity of zero because Built does not gain from Water during the time interval. The intensity of gain of Built in the extent of Non-Built at 1971 is 13% computed as the ratio of to 41 ha to 314 ha. The TOC curve communicates how Built's gain is more intensive on categories that come earlier in the sequence of Barren then Forest then Water. Each triangle that connects the TOC's segments indicates a threshold. The number of categories dictates the number of thresholds. The threshold label next to each triangle on the TOC curve indicates the additional land

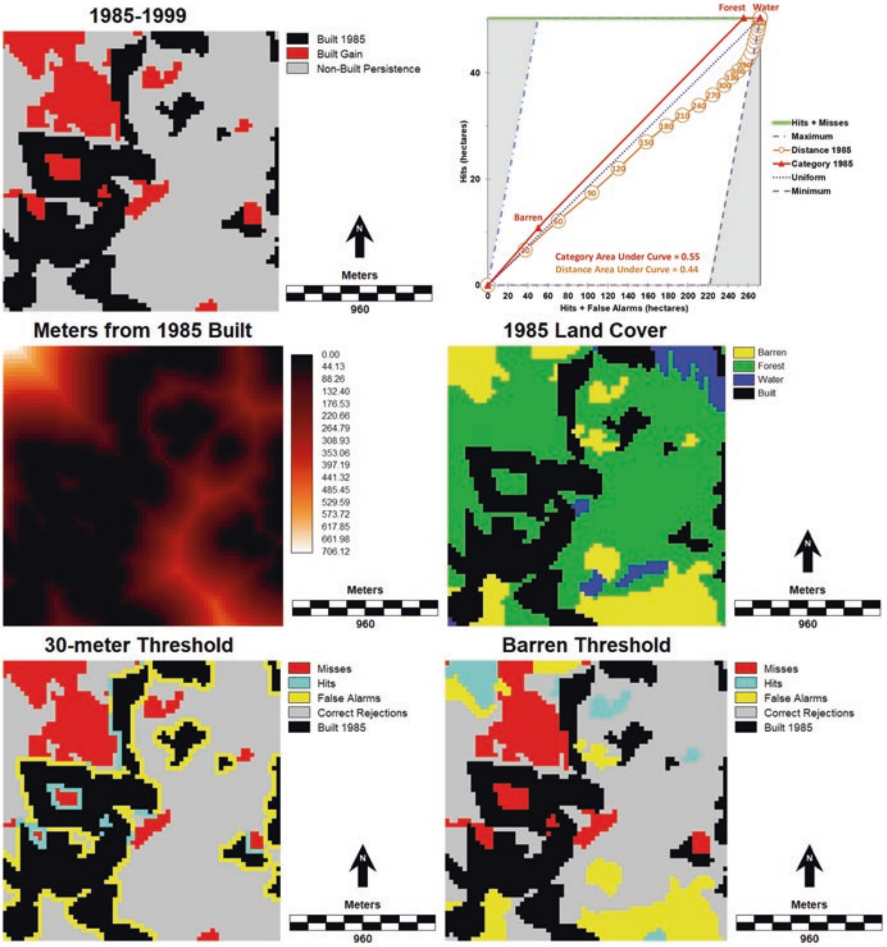


Fig. 3.2 Application of Total Operating Characteristic to describe Built's gain during 1985–1999

cover category that the sequence includes. The TOC curve begins at the origin of the TOC plot. The thresholds progress in sequence with priority ranking for the categories that have greater intensity of Built's gain. The threshold at Barren shows that the sum of Hits and False Alarms is 66 ha, which is greater than the 41 ha of Built's gain. The map in the lower right of Fig. 3.1 shows the Hits, False Alarms, Misses, and Correct Rejections for the threshold at Barren. The Hits and Misses are places where Built gained during 1971–1985. The Hits and False Alarms are places that are Barren at 1971. The subsequent threshold combines Barren and Forest to form the sum of Hits and False Alarms. This threshold labeled Forest is the point where the TOC curve meets the upper bound, which indicates that all of Built's gain derives from either Barren or Forest. The last threshold adds Water to the sequence of categories. Each pair of thresholds creates a bin. The slope of the TOC segment between thresholds is the intensity of Hits in the segment's bin. Specifically, a segment's slope is the increase in Hits as a proportion of the increase in the sum of Hits and False Alarms. A steeper slope of the segment between two thresholds indicates that Hits are more intensive in the bin that the thresholds create. The segment between the origin and the threshold at Barren has the steepest slope, which reflects the fact that Barren experiences the greatest intensity of Built's gain. Forest contains the next greatest intensity of Built's gain from the land cover categories. Water contains none of Built's gain, thus the slope is zero for the last segment in the upper right corner of the TOC. If the slope of a segment is greater than the slope of the uniform line, then the category that forms the segment has intensity greater than the intensity in the Non-Built at 1971. For example, Fig. 3.1 shows that the intensity for Barren is greater than uniform. The segment for Forest is nearly parallel to the uniform line, which indicates the intensity in Forest is nearly identical to the intensity in the Non-Built at 1971. The slope for Water is zero, which indicates that none of the Water at 1971 transitions to Built during 1971–1985. The TOC curve is above the uniform line by design for this ranked categorical variable, because the intensities determined the sequence of the ranking from Barren to Forest to Water. If all of the categories at 1971 had an equal intensity of Built's gain, then the TOC curve would be identical to the uniform line, in which case AUC would be 0.5. The AUC of 0.61 for the land category variable is greater than 0.5, which indicates that the Built's gain tends to occur more intensively on the categories that come earlier in the sequence of thresholds. The AUC value of 0.61 is closer to 0.5 than to 1, which indicates that the relationship between Built's gain and the ranked categories at 1971 are closer to uniform than to perfect. The AUC of 0.61 for land cover is farther from 0.5 than the AUC of 0.60 for distance, which suggests that Built's gain relationship with land cover is overall stronger than with distance. The AUC summarizes the direction and strength of the relationship, but AUC fails to reveal the details of the shape of the TOC curve. The TOC curve for land cover is sometimes above and sometimes below the TOC curve for distance. Therefore, the relationship between the gain of Built with land cover is not consistently stronger than with distance. Thus, the difference between the AUC values of 0.60 and 0.61 is not particularly meaningful because the AUC fails to show details of the TOC curve. This illustrates a limitation of the AUC and the importance of plotting more than one variable in the same TOC space.

Figure 3.2 describes Built's gain during 1985–1999 using the same approach as Fig. 3.1. The left side of Fig. 3.2 shows distance from Built at 1985. The TOC uses the same sequence of thresholds as the analysis during 1971–1985, meaning thresholds at increments of 30 m in a sequence that prioritizes smaller distances. The threshold at 30 m is the threshold where the quantity of the union of Hits and False Alarms is most similar to the quantity of the union of Hits and Misses, as the upper left corner of the TOC space indicates. The map in the lower left of Fig. 3.2 shows the Hits, False Alarms, Misses, and Correct Rejections at a distance threshold of 30 m. The Hits and False Alarms are within one 30-m pixel of the Built at 1985. Every point on the TOC curve for distance separates the extent into two parts: a part nearer the Built at 1985 and a part farther from Built at 1985. Every such point is below the uniform line, which indicates that Built's gain is less intensive in the part nearer to Built at 1985. Each pair of consecutive thresholds forms a straight segment of the TOC curve. The slopes of the segments are flatter than uniform from the origin of the TOC space up to the threshold of 300 m, which indicates that Built's gain is less intensive than uniform between consecutive thresholds that are less than or equal to 300 m from Built at 1985. The steepest segments of the TOC curve are in the upper right corner of the TOC space, which indicates that Built's gain is most intensive between consecutive thresholds that are farthest from Built at 1985. The AUC of 0.44 is less than 0.5, which indicates that Built's gain is overall less intensive between thresholds nearer to Built at 1985. The AUC of 0.44 is closer to 0.5 than to 0, which indicates that the strength of the relationship is closer to uniform than to perfect.

The right side of Fig. 3.2 analyzes the relationship Built's gain during 1985–1999 and land categories at 1985. The ranked sequence of categories is the same as for the first time interval, meaning Barren, Forest, and Water. The horizontal coordinate of Barren's point on the TOC curve is slightly to the right of the horizontal coordinate for the upper left corner of the TOC's parallelogram because Barren occupies slightly more pixels than the number of pixels of Built's gain. The lower right of Fig. 3.2 shows the map for the threshold at Barren. The map shows the spatial arrangement of Hits, False Alarms, Misses, and Correct Rejections, while the TOC does not indicate spatial arrangement. The slopes of the first two segments of the TOC are steeper than the uniform line, which indicates that Built's gain is more intensive in both Barren and Forest than in the Non-Built at 1985. The slope of the last segment is zero, which indicates that Built does not gain from Water. The AUC is 0.55 for land cover, compared to the AUC of 0.44 for distance. The AUC for land cover deviates 0.05 from 0.5 while the AUC for distance deviates 0.06 from 0.5. The smaller deviation for land cover indicates an overall weaker relationship than the larger deviation for distance. The AUC summarizes the strength of the relationship across all thresholds, which fails to reveal the details of the TOC's shape and number of thresholds. The TOC curve shows that the thresholds for land cover are fewer than the thresholds for distance. Furthermore, the thresholds for land cover do not have the same horizontal coordinates as the thresholds for distance. This illustrates how a comparison between TOC curves is more informative than comparison between AUC values. One must consider how the slope between thresholds compares to the uniform slope, and how far each threshold is above or below the uniform line. TOC curves that are farther from the uniform line indicate a stronger monotonic relationship with the ranked **X** variable.

Comparison between Figs. 3.1 and 3.2 shows the difference between the first and second time intervals. The TOC curve for distance is above the uniform line during the first time interval and below the uniform line during the second time interval. This indicates that Built's gain during the first time interval is more intensive nearer to Built while Built's gain during the second time interval is less intensive nearer to Built. The AUC supports this conclusion as the AUC of 0.60 during the first time interval is greater than 0.5 while the AUC of 0.44 during the second time interval is less than 0.5. Thus, the relationship between distance and Built's gain during the first time interval is opposite the relationship during the second time interval. Temporal non-stationarity is the phrase that describes a relationship that is not consistent from one time interval to a subsequent time interval. The AUC of 0.60 during the first time interval deviates more from 0.5 than the AUC of 0.44 during the second time interval, thus the strength of the relationship between distance and Built's gain is stronger during the first time interval than during the second time interval. If the relationship during a time interval matches the relationship during another time interval, then we say the relationship demonstrates temporal stationarity. However, there are many ways to characterize a relationship. For this case study, the first time interval is stationary with the second time interval in terms of the relationship between the intensity of Built's gain and the sequence of land cover categories. Barren, Forest, then Water is the sequence of categories in terms of intensity of Built's gain during both time intervals. The AUC concerning land cover during the first time interval is 0.61, which farther from 0.5 than the AUC of 0.55 during the second time interval, thus the strength of the relationship between land cover and Built's gain is overall stronger during the first time interval than during the second time interval.

Some readers might ask how severely two AUC values must deviate to qualify as an important deviation. That is a good question, and the answer depends on the definition of important. A good exercise for any scientist is to consider how to define important. The definition should depend on the goals of the analysis. The goals of this chapter's TOC curves are to describe the relationship between the gain of Built and two variables across two time intervals. The TOC curves show that the relationships are more subtle than a single metric such as AUC can communicate clearly. Thus, the AUC might be uninformative and potentially misleading for some questions. Nevertheless, it can be helpful to have a summary metric such as AUC for broad comparisons. The AUC is 0.61 for land cover and is 0.60 for distance during the first time interval. Thus, readers might wonder whether we should consider the deviation of 0.01 as important, and whether either AUC deviates in an important manner from 0.50. To address this concern, it is necessary to examine whether various types of uncertainty could explain the possible variance in AUC values. Hypothesis tests and inferential statistics are not applicable to compare AUC values for this case study, because the data derive from a census of all the pixels in the extent, not from a sample. However, other sources of uncertainty exist. For example, there is uncertainty concerning data quality because the underlying data might contain errors that could influence the AUC values. There might be so much error in the maps that it would be misleading to place any importance on a deviation of 0.01 or 0.1 between AUC values when considering the difference between the data and the true landscape. However, we do not know the errors in the maps because the maps derive from the best available information, thus we cannot measure the uncertainty due to data quality.

We can measure uncertainty concerning AUC when the TOC algorithm fails to select a threshold at every unique value in the **X** variable. The algorithm for the TOC in Figs. 3.1 and 3.2 selected thresholds at 30-m increments of the distance variable. However, some observations have various distances in the bins that the 30-m increments form. A refined algorithm could have selected a threshold at each unique distance, which would generate a TOC curve for all possible thresholds. An algorithm eliminates uncertainty due to threshold selection when the algorithm selects a threshold at every unique value of the **X** variable, which is an option for the algorithm of the TOC package in the computer language R (Pontius Jr et al. 2015) and for the TOC Curve Generator (Liu 2020). This option requires substantial computing power when the number of observations is large. I used that option to compute the AUC for each of the time intervals in this chapter. During 1971–1985, the AUC that derives from thresholds at all unique distances matches within two decimal places the AUC that derives from 30-m thresholds, i.e. both AUC values for distance are 0.60. Thus, threshold selection at 30-meter increments for distance does not influence the conclusion that the AUC of 0.60 for distance is smaller than the AUC of 0.61 for land cover. Furthermore, threshold selection does not influence the conclusion that the AUC of 0.60 for distance is greater than 0.5. During the second time interval, the AUC that derives from all possible thresholds of distance is 0.43 while the AUC that derives from the 30-m thresholds is 0.44. Thus, threshold selection at 30-m increments for distance does not influence the conclusion that the strength of the relationship with distance is stronger than with land cover because both 0.44 and 0.43 deviate more from 0.5 than does 0.55, which is the AUC for land cover. Furthermore, threshold selection does not influence the conclusion that the strength of the relationship with distance during the second time interval is weaker than during the first time interval because both 0.44 and 0.43 deviate less from 0.5 than does 0.60, which is the AUC for distance during the first time interval. Pontius and Parmentier (2014) gives more details concerning how threshold selection can influence the uncertainty in AUC values.

3.2 Discussion Questions

1. How can the TOC curve use a categorical variable as the **X** variable?
2. What is the interpretation when the entire TOC curve is either above or below the uniform line?
3. How would you interpret the TOC curve when it is above the diagonal uniform line at some thresholds and below the uniform line at other thresholds?
4. What do you learn from comparing the slope of a line segment on the TOC curve to the slope of the uniform line?
5. What is the interpretation of the point where the TOC curve meets the horizontal portion of the maximum upper bound or the non-horizontal portion of the minimum bound?
6. What are the advantages and disadvantages of using the AUC as a metric to summarize the TOC?
7. What is the maximum number of thresholds on a TOC curve for a particular **X** variable?

8. How can you determine whether the difference between two AUC values is important?

References

Alves, J., Silva, J., Grifo, E., Resende, C., & Sousa, I. (2019). Wearable embedded intelligence for detection of falls independently of on-body location. *Sensors, 19*, 2426. https://doi.org/10.3390/s19112426.

Chakraborti, S., Das, D. N., Mondal, B., Shafizadeh-Moghadam, H., & Feng, Y. (2018). A neural network and landscape metrics to propose a flexible urban growth boundary: A case study. *Ecological Indicators, 93*, 952–965. https://doi.org/10.1016/j.ecolind.2018.05.036.

Cushman, S. A., Macdonald, E. A., Landguth, E. L., Malhi, Y., & Macdonald, D. W. (2017). Multiple-scale prediction of forest loss risk across Borneo. *Landscape Ecology, 32*, 1581–1598. https://doi.org/10.1007/s10980-017-0520-0.

Dustin, D. L., & Jacobson, P. C. (2015). Predicting the extent of lakeshore development using GIS datasets. *Lake and Reservoir Management, 31*, 169–179. https://doi.org/10.1080/10402381.2015.1053010.

Kamusoko, C., & Gamba, J. (2015). Simulating urban growth using a random forest-cellular automata (RF-CA) model. *ISPRS International Journal of Geo-Information, 4*, 447–470. https://doi.org/10.3390/ijgi4020447.

Liu, Z. (2020). *TOC Curve Generator*. https://lazygis.github.io/projects/TOCCurveGenerator

Naghibi, F., & Delavar, M. (2016). Discovery of transition rules for cellular automata using artificial bee colony and particle swarm optimization algorithms in urban growth modeling. *ISPRS International Journal of Geo-Information, 5*, 241. https://doi.org/10.3390/ijgi5120241.

Pontius Jr, R. G., & Parmentier, B. (2014). Recommendations for using the relative operating characteristic (ROC). *Landscape Ecology, 29*, 367–382. https://doi.org/10.1007/s10980-013-9984-8.

Pontius Jr, R. G., & Si, K. (2014). The total operating characteristic to measure diagnostic ability for multiple thresholds. *International Journal of Geographical Information Science, 28*, 570–583. https://doi.org/10.1080/13658816.2013.862623.

Pontius Jr, R. G., Santacruz, A., Tayyebi, A., & Parmentier, B. (2015). *TOC: Total Operating Characteristic curve and ROC curve*. https://cran.r-project.org/web/packages/TOC

Shafizadeh-Moghadam, H., Asghari, A., Tayyebi, A., & Taleai, M. (2017a). Coupling machine learning, tree-based and statistical models with cellular automata to simulate urban growth. *Computers, Environment and Urban Systems, 64*, 297–308. https://doi.org/10.1016/j.compenvurbsys.2017.04.002.

Shafizadeh-Moghadam, H., Tayyebi, A., Ahmadlou, M., Delavar, M. R., & Hasanlou, M. (2017b). Integration of genetic algorithm and multiple kernel support vector regression for modeling urban growth. *Computers, Environment and Urban Systems, 65*, 28–40. https://doi.org/10.1016/j.compenvurbsys.2017.04.011.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better DECISIONS through SCIENCE. *Scientific American, 283*, 82–87. https://www.scientificamerican.com/article/better-decisions-through-science/.

Chapter 4
Categorical Variable Versus Categorical Variable

Abstract This chapter gives methods to compare variables **X** and **Y** when both indicate the same group of categories. The analysis' foundation is an extended square contingency table where the sequence of categories for **X** in the rows is identical to the sequence of categories for **Y** in the columns. The table's diagonal entries show agreement while the off-diagonal entries show difference. Equations express the sizes and intensities of the differences as the sum of three components: Quantity, Exchange, and Shift. Equations express also the intensities of the entries and the categories in the table's rows and columns. Interpretation of the results for the entries and categories depends on whether the sums of the rows can influence the sums of the columns, vice-versa, or neither. Relevant software includes the PontiusMatrix42.xlsx spreadsheet available at www.clarku.edu/~rpontius (Pontius Jr 2020), the pontiPy python code available at https://github.com/verma-priyanka/pontiPy (Ahn and Verma 2021), and the diffeR package available at https://cran.r-project.org/web/packages/diffeR/index.html (Pontius Jr and Santacruz 2015).

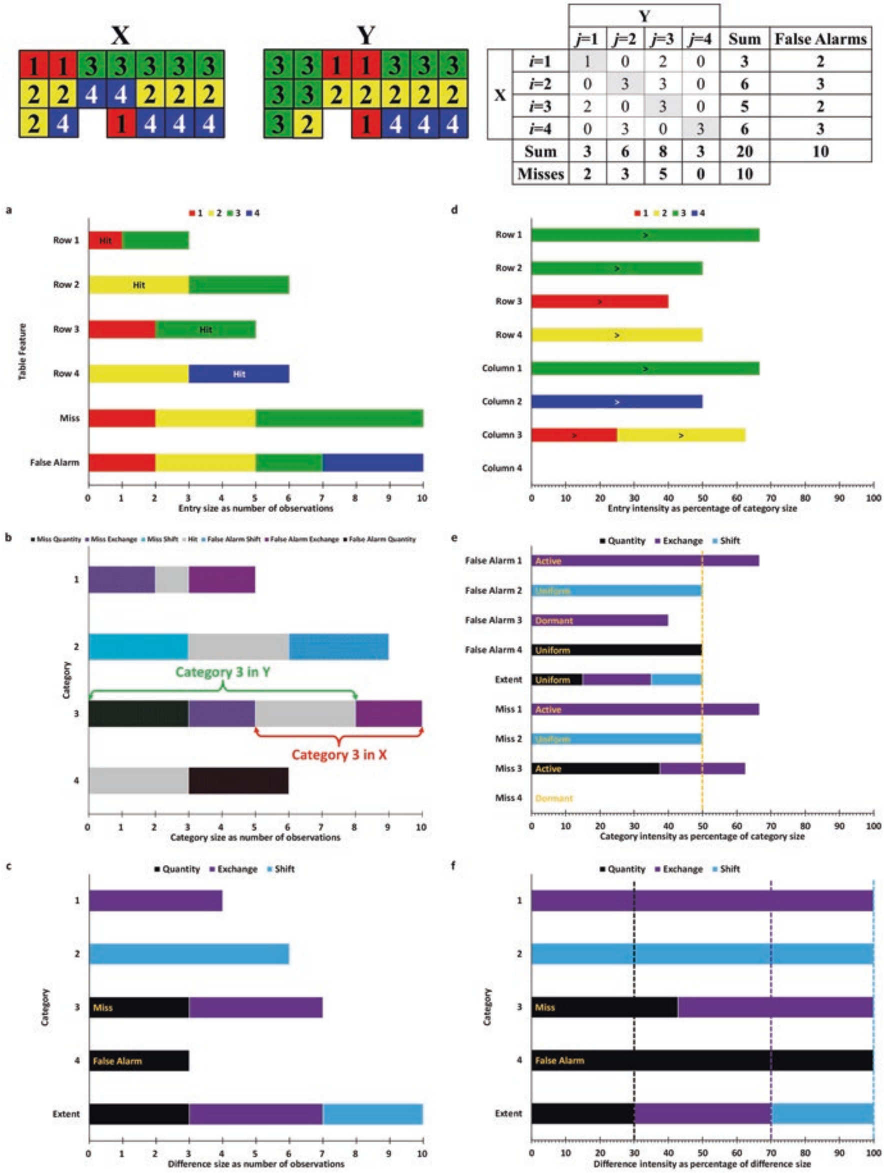
Keywords Contingency table · Category · Exchange · Intensity · PontiusMatrix · Shift · Quantity

4.1 Text

Categorical is the name of a type of variable that classifies each observation as a category. Literature refers to this type of variable also as Discrete or Nominal. This chapter shows how to compare two categorical variables when both variables use the same group of categories to classify the observations. This chapter analyzes each category in the manner that Chap. 1 analyzes the Presence category.

Figure 4.1 illustrates the concepts. The upper left of Fig. 4.1 shows the example data in the form of square pixels. The number of observations is 20. The number of categories is four, named 1, 2, 3 and 4. An extended square contingency table in the upper right of Fig. 4.1 summarizes the association between **X** and **Y**. The categories of **X** are in the table's rows while the same sequence of categories for **Y** are in the table's columns. The table is square in the respect that the number of columns equals the number of rows. The table's diagonal entries give the number of observations for

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1_4



X

1	1	3	3	3	3	3
2	2	4	4	2	2	2
2	4		1	4	4	4

Y

3	3	1	1	3	3	3
3	3	2	2	2	2	2
3	2		1	4	4	4

		Y					
		j=1	*j*=2	*j*=3	*j*=4	Sum	False Alarms
X	*i*=1	1	0	2	0	3	2
	i=2	0	3	3	0	6	3
	i=3	2	0	3	0	5	2
	i=4	0	3	0	3	6	3
	Sum	3	6	8	3	20	10
	Misses	2	3	5	0	10	

Fig. 4.1 Example to compare two variables that show the same group of categories

which the **X** category matches the **Y** category, in which case the observations are Hits for the category. The off-diagonal entries give the number of observations for which the **X** category differs from the **Y** category. Each off-diagonal entry is a False Alarm for the **X** category and a Miss for the **Y** category. The sums in the right margin give the size of each category for **X** while the sums in the bottom margin give the size of each category for **Y**. Figure 4.1 extends the table to give an additional column at the right labeled False Alarms, which gives the sum of the off-diagonal entries in each row. The extended table gives also a row at the bottom labeled Misses, which gives the sum of the off-diagonal entries in each column. This accounting framework assures for all cases that the sum of False Alarms equals the sum of Misses, which equals the sum of the off-diagonal entries. The graphs in Fig. 4.1 are results that derive from this chapter's equations. The graphs on the left side of Fig. 4.1 labeled a–c give results in terms of sizes. The graphs on the right side of Fig. 4.1 labeled d–f give corresponding results in terms of intensities. The sizes in Fig. 4.1a produce the intensities in Fig. 4.1d. The sizes in Fig. 4.1b produce the intensities in Fig. 4.1e. The sizes in Fig. 4.1c produce the intensities in Fig. 4.1f. Table 4.1 gives the mathematical notation for the equations that compute the results in Fig. 4.1.

The contingency table's entries along with Eqs. 4.1 and 4.2 generate Fig. 4.1a and the graph on the EntrySize sheet in PontiusMatrix42.xlsx. The entire length of each category's bar is the size of each category *i* in **X**. The segments in each category's bar are the sizes of the categories in **Y**, which the legend at the top of Fig. 4.1a denotes. Each diagonal entry in the table has a label of Hit in Fig. 4.1a. Equation 4.1 gives the size of the False Alarm for category *k*. Equation 4.2 gives the size of the Miss for category *k*. The sum of False Alarms over all categories equals the sum of Misses over all categories, which equals the Difference for the extent as Fig. 4.1a shows by the entire length of its bottom two bars.

$$F_k = \left(\sum_{j=1}^{J} N_{kj} \right) - N_{kk} \tag{4.1}$$

$$M_k = \left(\sum_{i=1}^{J} N_{ik} \right) - N_{kk} \tag{4.2}$$

Equations 4.3, 4.4, 4.5, 4.6 and 4.7 generate Fig. 4.1b and the graph on the CategorySize sheet in PontiusMatrix42.xlsx. Figure 4.1b shows a horizontal Venn diagram for each category in the same manner that Chap. 1 showed a horizontal Venn diagram for the Presence category. The braces show how category 3 in **X** partially intersects category 3 in **Y** to help the reader envision that each segmented bar is a Venn diagram, where the intersection is the size of Hits for a category. Hits for category *k* are observations that are category *k* in both **X** and **Y**. Equation 4.3 shows that the size of Hits for category *k* is the table's diagonal entry. False Alarms for category *k* appear to the right of Hits in the category's Venn diagram, just as False Alarms appear at the right in the extended table. Misses for category *k* appear to the left of Hits in the category's Venn diagram. If the sum of observations of *k* in **X** does not equal the sum of observations of *k* in **Y**, then the size of False Alarms for *k* does not equal the size of Misses for *k*, in which case either False Alarms or Misses has a positive Quantity component. Equation 4.4 gives the size of the False Alarm Quantity component for category *k*. Equation 4.5 gives the size of the Miss Quantity component for category *k*. In the example data, categories 1 and 2 have Quantity

Table 4.1 Notation to compare two variables that show the same categorical phenomenon

Notation	Meaning
D	Difference for the extent
D_k	Difference for category k
De	Difference Exchange for the extent
De_k	Difference Exchange for category k
Ds	Difference Shift for the extent
Ds_k	Difference Shift for category k
Dq	Difference Quantity for the extent
Dq_k	Difference Quantity for category k
F_i	False Alarms for category i
F_k	False Alarms for category k
Fe_k	False Alarm Exchange for category k
Fs_k	False Alarm Shift for category k
Fq_k	False Alarm Quantity for category k
H_k	Hits for category k
M_j	Misses for category j
M_k	Misses for category k
Me_k	Miss Exchange for category k
Ms_k	Miss Shift for category k
Mq_k	Miss Quantity for category k
i	Index for a category where $i = 1, 2, \ldots J$
j	Index for a category where $j = 1, 2, \ldots J$
J	Number of categories
k	Index for a category where $k = 1, 2, \ldots J$
N_{ii}	Number of observations in both row i and column i
N_{ij}	Number of observations in both row i and column j
N_{ik}	Number of observations in both row i and column k
N_{jj}	Number of observations in both row j and column j
N_{jk}	Number of observations in both row j and column k
N_{ki}	Number of observations in both row k and column i
N_{kj}	Number of observations in both row k and column j
N_{kk}	Number of observations in both row k and column k

components of zero, while category 3 has a Miss Quantity component of three whereas category 4 has a False Alarm Quantity component of three. If category k has both False Alarms and Misses, then category k has at least one of the components called Exchange or Shift. Exchange forms where each observation in row k and column j is paired with an observation in row j and column k. The number of pairs between k and j is the smaller of N_{kj} and N_{jk}. Equation 4.6 sums the pairs for category k over all categories for which $j \neq k$. The False Alarm Exchange component in the row for category k equals the Miss Exchange component in the column for category k. For category 1 in the example, the False Alarm Exchange component and the Miss Exchange component are both two, as category 1 exchanges with category 3. If the number of categories is larger than two, then it is possible to have a component called Shift. Shift is the difference that is neither Quantity nor Exchange. Equation 4.7 computes the False Alarm Shift component as the False Alarm minus both Quantity and Exchange. Equation 4.7 shows that the False Alarm Shift in row *k* equals the Miss Shift in column *k*. Shift for category *k* is positive when there exists categories *i* and *j* such that $N_{kj} > N_{jk}$ and $N_{ki} < N_{ik}$. For example, the example shows $N_{23} = 3 > N_{32} = 0$ while $N_{24} = 0 < N_{42} = 3$ thus both False Alarm Shift and Miss Shift are three for category 2.

$$H_k = N_{kk} \tag{4.3}$$

$$Fq_k = \mathrm{MAXIMUM}\left(0, F_k - M_k\right) \tag{4.4}$$

$$Mq_k = \mathrm{MAXIMUM}\left(0, M_k - F_k\right) \tag{4.5}$$

$$Fe_k = Me_k = \left[\sum_{j=1}^{J} \mathrm{MINIMUM}\left(N_{kj}, N_{jk}\right)\right] - N_{kk} \tag{4.6}$$

$$Fs_k = F_k - Fq_k - Fe_k = Ms_k = M_k - Mq_k - Me_k \tag{4.7}$$

$$Dq_k = Fq_k + Mq_k = \left|F_k - M_k\right| \tag{4.8}$$

$$De_k = Fe_k + Me_k \tag{4.9}$$

$$Ds_k = Fs_k + Ms_k \tag{4.10}$$

$$D_k = F_k + M_k = Dq_k + De_k + Ds_k \tag{4.11}$$

$$Dq = \sum_{k=1}^{J} Fq_k = \sum_{k=1}^{J} Mq_k \tag{4.12}$$

$$De = \sum_{k=1}^{J} Fe_k = \sum_{k=1}^{J} Me_k \tag{4.13}$$

$$Ds = \sum_{k=1}^{J} Fs_k = \sum_{k=1}^{J} Ms_k \tag{4.14}$$

$$D = Dq + De + Ds = \sum_{k=1}^{J} F_k = \sum_{k=1}^{J} M_k \tag{4.15}$$

Equations 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14 and 4.15 generate Fig. 4.1c and the graph on the DifferenceSize sheet in PontiusMatrix42.xlsx. Figure 4.1c shows the difference for each category, which is the sum of the category's False Alarms in its row and the category's Misses in its column. Equations 4.8, 4.9 and 4.10 compute each category's three components: Quantity, Exchange, and Shift (Pontius Jr and Santacruz 2014, 2015). A category's component is the sum of the category's False Alarms and Misses for the respective component. Equation 4.11 shows that a category's difference is the sum of the category's False Alarms and Misses, which is also equal to the sum of the category's three components. Equations 4.12, 4.13 and 4.14 show the components for the extent, where the extent means summed over all categories. Equation 4.15 shows that the difference for the extent is equal to three sums: the sum of the three components for the extent, the sum of False Alarms over all categories, and the sum of Misses over all categories. The Quantity component for each category is by definition zero or positive. A positive component does not indicate whether False Alarms are greater than Misses or whether Misses are greater than False Alarms. Therefore, Fig. 4.1c gives a label on the Quantity component to denote the larger of the category's False Alarms or Misses. For example, the Quantity component for category 3 is positive because its Misses are larger than its False Alarms, whereas the Quantity component for category 4 is positive because its False Alarms are larger than its Misses. The Quantity component for the extent does not have such a label because it derives from False Alarms for some categories and Misses for other categories. Equations 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14 and 4.15 compute differences in terms of size, whereas the remaining equations in this chapter compute differences in terms of intensities that range from 0% to 100%.

Equations 4.16, 4.17, 4.18 and 4.19 generate Fig. 4.1d and the graph on the EntryIntensity sheet of PontiusMatrix42.xlsx. Equation 4.16 is a percentage where the numerator is the size of the False Alarms of k and the denominator is the size of category k in $\mathbf{X}$. This percentage is also known as commission error intensity for applications to error assessment and as loss intensity for applications to change analysis. Equation 4.16 is the sum of $J-1$ parts, where Eq. 4.17 gives the part that derives from category j in $\mathbf{Y}$. Equation 4.18 is a percentage where the numerator contains the size of the Misses of k and the denominator is the size of category k in $\mathbf{Y}$. This percentage is also known as omission error intensity for applications to error assessment and as gain intensity for applications to change analysis. Equation 4.18 is the sum of $J-1$ parts, where Eq. 4.19 gives the part that derives from category i in $\mathbf{X}$. The end of this chapter explains the labels of the greater than symbols on the segments in Fig. 4.1d.

$$\text{Category Intensity of False Alarm for } k = \frac{F_k\,100\%}{\text{size of } k \text{ in } \mathbf{X}} = \frac{F_k\,100\%}{F_k + H_k} \quad (4.16)$$

$$\text{Entry Intensity of False Alarm for } k \text{ in column } j = \frac{N_{kj}\,100\%}{F_k + H_k} \text{ for } j \neq k \quad (4.17)$$

$$\text{Category Intensity of Miss for } k = \frac{M_k\,100\%}{\text{size of } k \text{ in } \mathbf{Y}} = \frac{M_k\,100\%}{M_k + H_k} \quad (4.18)$$

$$\text{Entry Intensity of Miss for } k \text{ in row } i = \frac{N_{ik}\,100\%}{M_k + H_k} \text{ for } i \neq k \quad (4.19)$$

Equations 4.20, 4.21, 4.22, 4.23, 4.24, 4.25, 4.26, 4.27, 4.28, 4.29, 4.30 and 4.31 generate Fig. 4.1e and the graph on the CategoryIntensity sheet in the PontiusMatrix42

Excel file (Pontius Jr 2019). The lengths of the bars in Fig. 4.1e match the lengths of the corresponding bars in Fig. 4.1d. Figure 4.1d shows how the categories contribute to the length of the entire intensity bar, whereas Fig. 4.1e shows how the components contribute to the length of the entire intensity bar. Equations 4.20, 4.21 and 4.22 give the False Alarm intensity for each of the three components by expressing each False Alarm component for *k* as a percentage of the size of category *k* in **X**, meaning in row *k* of the table. Equation 4.23 gives the False Alarm intensity for category *k*, which is the sum of the three component intensities for category *k*. Equations 4.24, 4.25 and 4.26 give the Miss intensity for each component by expressing the Miss for *k* as a percentage of the size of category *k* in **Y**, meaning in column *k* of the table. Equation 4.27 gives the Miss intensity for category *k*, which is the sum of the three component intensities for category *k*. Equations 4.28, 4.29 and 4.30 give the intensity for each component of difference in the extent by computing the size of the component as a percentage of the sum of all entries in the contingency table. Equation 4.31 gives the intensity of difference for the extent expressed as a percentage of the sum of all entries in the contingency table. The extent's difference intensity in Eq. 4.31 offers a helpful baseline to interpret the intensity of each category's False Alarm or Miss. If the intensity of the False Alarm or Miss of category *k* is less than the intensity of the extent's difference, then we say the intensity of the respective False Alarm or Miss of category *k* is dormant. If the intensity of the False Alarm or Miss of category *k* is equal to the intensity of the extent's difference, then we say the intensity of the respective False Alarm or Miss of category *k* is uniform. If the intensity of the False Alarm or Miss of category *k* is greater than the intensity of the extent's difference, then we say the intensity of the respective False Alarm or Miss of category *k* is active. For example, the labels on the bars in Fig. 4.1e show that category 2 is uniform in both its False Alarm and Miss, while category 3 is dormant in its False Alarm and active in its Miss.

$$\text{Quantity Intensity of False Alarm for } k = \frac{Fq_k\,100\%}{F_k + H_k} \tag{4.20}$$

$$\text{Exchange Intensity of False Alarm for } k = \frac{Fe_k\,100\%}{F_k + H_k} \tag{4.21}$$

$$\text{Shift Intensity of False Alarm for } k = \frac{Fs_k\,100\%}{F_k + H_k} \tag{4.22}$$

$$\text{Intensity of False Alarm for } k = \frac{F_k\,100\%}{\text{size of } k \text{ in } \mathbf{X}} = \frac{\left(Fq_k + Fe_k + Fs_k\right)100\%}{F_k + H_k} \tag{4.23}$$

$$\text{Quantity Intensity of Miss for } k = \frac{Mq_k\,100\%}{M_k + H_k} \tag{4.24}$$

$$\text{Exchange Intensity of Miss for } k = \frac{Me_k\,100\%}{M_k + H_k} \tag{4.25}$$

$$\text{Shift Intensity of Miss for } k = \frac{Ms_k\,100\%}{M_k + H_k} \tag{4.26}$$

$$\text{Intensity of Miss for } k = \frac{M_k\,100\%}{\text{size of } k \text{ in } \mathbf{Y}} = \frac{\left(Mq_k + Me_k + Ms_k\right)100\%}{M_k + H_k} \tag{4.27}$$

$$\text{Quantity Intensity of Difference in Extent} = \frac{Dq\,100\%}{\sum_{i=1}^{J}\sum_{j=1}^{J} N_{ij}} \tag{4.28}$$

$$\text{Exchange Intensity of Difference in Extent} = \frac{De\,100\%}{\sum_{i=1}^{J}\sum_{j=1}^{J} N_{ij}} \tag{4.29}$$

$$\text{Shift Intensity of Difference in Extent} = \frac{Ds\,100\%}{\sum_{i=1}^{J}\sum_{j=1}^{J} N_{ij}} \tag{4.30}$$

$$\text{Intensity of Difference in Extent} = \frac{D\,100\%}{\sum_{i=1}^{J}\sum_{j=1}^{J} N_{ij}} \tag{4.31}$$

$$\text{Quantity Intensity of Difference for } k = \frac{Dq_k\,100\%}{D_k} \tag{4.32}$$

$$\text{Exchange Intensity of Difference for } k = \frac{De_k\,100\%}{D_k} \tag{4.33}$$

$$\text{Shift Intensity of Difference for } k = \frac{Ds_k\,100\%}{D_k} \tag{4.34}$$

$$\text{Quantity Intensity in Difference} = \frac{Dq\,100\%}{D} \tag{4.35}$$

$$\text{Exchange Intensity in Difference} = \frac{De\,100\%}{D} \tag{4.36}$$

$$\text{Shift Intensity in Difference} = \frac{Ds\,100\%}{D} \tag{4.37}$$

Equations 4.32, 4.33, 4.34, 4.35, 4.36 and 4.37 generate Fig. 4.1f and the graph on the DifferenceIntensity sheet in PontiusMatrix42.xlsx. Equations 4.32, 4.33 and 4.34 compute each component's intensity for category *k* as a percentage of the size of the category's difference. Equations 4.35, 4.36 and 4.37 compute each component's intensity as a percentage of the extent's difference. Equations 4.32, 4.33, and 4.34 sum to 100%; similarly, Eqs. 4.35, 4.36 and 4.37 sum to 100%. Figure 4.1f allows the reader to characterize the intensity of each component for each category relative to the intensity of the respective component for the extent's difference. Figure 4.1f shows that difference for the extent consists of 30% Quantity, 40%

Exchange, and 30% Shift. The Quantity component for category 3 accounts for approximately 43% of the difference for category 3. This 43% is larger than the 30% of Quantity intensity for the extent's difference, which means that category 3 has a Quantity component that is more intensive than the extent's Quantity component. The label in the Quantity component denotes Miss, which indicates that the size of Miss is larger than the size of False Alarm for category 3. The Exchange component for category 3 accounts for approximately 57% of the difference for category 3. This 57% is larger than the 40% of Exchange intensity for extent's difference, which means that category 3 has an Exchange component that is more intensive than the extent's Exchange component.

It is tempting to envision that the numbers of observations N_{ij} for the central entries in the contingency table at the top of Fig. 4.1 cause the marginal sums, False Alarms, and Misses. However, the mathematical accounting does not imply causation. For example, the number of observations in the extent exists independently from how the observations are distributed among the N_{ij} for the central entries. It is helpful to envision that the sum of all observations constrains the number of observations in the table's entries. Furthermore, the particular application determines whether the entries N_{ij} influence the rows' marginal sums or the columns' marginal sums. We must consider three types of applications when interpreting the marginal sums, False Alarms, and Misses. Each type of application calls for a distinct method to determine the labels on the segments of the bars in Fig. 4.1a, d. The labels describe how each entry compares to a uniform distribution of entries in a row and in a column. This chapter's remaining equations specify the uniform distributions for each type of application. Each of the next three paragraphs describe each of the three types of applications, where Fig. 4.1 is the third type.

The first type applies where the categories in **Y** exist regardless of **X**, as in error assessment where the convention is that **Y** is the truth in the columns and **X** is the diagnosis in the rows (Shafizadeh-Moghadam et al. 2019). The truth can influence the diagnosis but the diagnosis cannot not influence the truth. In this case, the marginal sums at the bottom of the table exist regardless of the diagnosis. Each marginal sum at the bottom of each column is distributed among the entries above. The entries then influence the marginal sums at the right side of the table. Equations 4.38 and 4.39 apply to this first type of application where **Y** can influence **X** but **X** cannot not influence **Y**. Equation 4.38 gives the uniform size of the off-diagonal entries. If the Misses in column j were distributed with uniform size among the column's off-diagonal entries, then Eq. 4.38 would be the size of the uniform entry that is a Miss for category j in **Y** and a False Alarm for each of the $J-1$ categories that are not j in **X**. For each off-diagonal entry in column j, if N_{ij} is greater than the result from Eq. 4.38, then the segment for N_{ij} receives a label of > in Fig. 4.1a to denote that the empirical size is greater than the uniform size. If N_{ij} is equal to the result from Eq. 4.38, then the segment for N_{ij} receives a label of = in Fig. 4.1a. If N_{ij} is less than the result from Eq. 4.38, then the segment for N_{ij} does not receive a label in Fig. 4.1a; the reason for lack of a label is to reduce clutter and to avoid labels for entries that have zero size. Equation 4.39 gives the uniform entry intensity. If the False Alarms in row i were distributed with uniform intensity among the row's off-diagonal entries, then Eq. 4.39 would be the uniform intensity that is a False Alarm for category i and a Miss for each of the $J - 1$ off-diagonal entries in row i. Equation 4.39 expresses how the False Alarms for i must derive from the categories that are not i in **Y**. For each off-diagonal entry in row i, if the empirical intensity from Eq. 4.19 is greater than the uniform intensity from Eq. 4.39, then the segment for the off-diagonal entry receives a label of > in the row's bar at the top of Fig. 4.1d to denote that the empirical intensity is greater than the uniform intensity. For each off-diagonal entry in row i, if the empirical intensity from Eq. 4.19 equals the uniform intensity from Eq. 4.39, then the segment for the off-diagonal entry receives a label of = in the row's bar at the top of Fig. 4.1d. If the empirical intensity from Eq. 4.19 is less than the uniform intensity from Eq. 4.39, then the segment for the off-diagonal entry does not receive a label in the bars at the top of Fig. 4.1d. Chapter 5 gives an example for this type of application to error assessment.

$$\text{Uniform Entry Size in column } j = \frac{M_j}{J-1} \tag{4.38}$$

$$\text{Uniform Entry Intensity in row } i = \frac{F_i 100\%}{\text{size of not } i \text{ in } \mathbf{Y}} = \frac{\left[\left(\sum_{j=1}^{J} N_{ij}\right) - N_{ii}\right] 100\%}{\left(\sum_{k=1}^{J}\sum_{j=1}^{J} N_{kj}\right) - \sum_{k=1}^{J} N_{ki}} \tag{4.39}$$

The second type applies where the categories in **X** exist regardless of **Y**, as in change analysis where the convention is that **X** is the start time in the rows and **Y** is the end time in the columns (Pontius Jr et al. 2017). The start time can influence the end time but the end time cannot influence the start time. For change analysis, False Alarms in row i are losses from category i, while Misses in column j are gains to category j. In this case, the marginal sums at the right side of the table are the start sizes, which exist regardless of the change. Each marginal sum at the right side of each row is distributed among the entries to the left. The entries then influence the marginal sums at the bottom of the table, which are the end sizes. Equations 4.40 and 4.41 apply to this second type of application where **X** can influence **Y** but **Y** cannot influence **X**. Equation 4.40 gives the uniform size of the off-diagonal entries. If the False Alarms in row i were distributed with uniform size among the row's off-diagonal entries, then Eq. 4.40 would be the size of the uniform entry that is a Miss for category j in **Y** and a False Alarm for each of the $J - 1$ categories that are not j in **X**. For each off-diagonal entry in row i, if N_{ij} is greater than the result from Eq. 4.40, then the segment for N_{ij} receives a label of > in Fig. 4.1a to denote that the empirical size is greater than the uniform size. If N_{ij} is equal to the result from Eq. 4.40, then the segment for N_{ij} receives a label of = in Fig. 4.1a. Equation 4.41 gives the uniform intensity. If the Misses in column j were distributed with uniform intensity among the row's off-diagonal entries, then Eq. 4.41 would be the uniform intensity that is a False Alarm for category i and a Miss for each of the $J - 1$ off-diagonal entries in column j. Equation 4.41 expresses how the Misses for j must derive from the categories that are not j in **X**. For each off-diagonal entry in column j, if the empirical intensity from Eq. 4.17 is greater than the uniform intensity from

Eq. 4.41, then the segment for the off-diagonal entry receives a label of > in the column’s bar at the bottom of Fig. 4.1d to denote that the empirical intensity is greater than the uniform intensity. In this case, the language of Intensity Analysis for change analysis says that the gain of category *j* targets category *i* (Aldwaik and Pontius Jr 2012, 2013; Enaruvbe and Pontius Jr 2015; Pontius Jr et al. 2013; Quan et al. 2020; Shafizadeh-Moghadam et al. 2019). For each off-diagonal entry in column *j*, if the empirical intensity from Eq. 4.17 equals the uniform intensity from Eq. 4.41, then the segment for the off-diagonal entry receives a label of = in the column’s bar at the bottom of Fig. 4.1d. Chapter 7 gives an example for this type of application to change analysis.

$$\text{Uniform Entry Size in row}\,i = \frac{F_i}{J-1} \tag{4.40}$$

$$\text{Uniform Entry Intensity in column}\,j = \frac{M_j\,100\%}{\text{size of not}\,j\,\text{in}\,\mathbf{X}} = \frac{\left[\left(\sum_{i=1}^{J} N_{ij}\right) - N_{jj}\right]100\%}{\left(\sum_{i=1}^{J}\sum_{k=1}^{J} N_{ik}\right) - \sum_{k=1}^{J} N_{jk}} \tag{4.41}$$

The third type applies when **X** is a diagnosis and **Y** is a different diagnosis, while the truth is not considered and might be unknown. This chapter’s example in Fig. 4.1 illustrates this third type, because this chapter’s example does not distinguish **X** from **Y** conceptually. A False Alarm of category *k* is where **X** diagnoses *k* while **Y** diagnoses a different category. A Miss of category *k* is where **Y** diagnoses *k* while **X** diagnoses a different category. In this case, both marginal sums exist independently of the association between **X** and **Y** whereas the entries N_{ij} indicate how **X** is associated with **Y**. Equations 4.39 and 4.41 apply to this third type of application for which neither **X** nor **Y** influence each other. If the result from Eq. 4.19 is greater than the result from Eq. 4.39, then the empirical False Alarm intensity receives a label of > in the relevant segment in the rows’ bars at the top of Fig. 4.1d. If the result from Eq. 4.17 is greater than the result from Eq. 4.41, then the empirical Miss intensity receives a label of > in the relevant segment in the columns’ bars at the bottom of Fig. 4.1d. The segments of Fig. 4.1a lack labels because Eqs. 4.38 and 4.40 lack helpful insight for this third type of application. It is possible that N_{ij} is simultaneously greater than Eq. 4.38 and less than Eq. 4.40. It is also possible that N_{ij} is simultaneously less than Eq. 4.38 and greater than Eq. 4.40.

Free software packages compute this chapters equations. The spreadsheet PontiusMatrix42.xlsx performs the calculations (Pontius Jr 2020). The user types the contingency table into the Input sheet, and then the spreadsheet computes the numerical results and presents them in graphical form. PontiusMatrix42.xlsx is available for free from www.clarku.edu/~rpontius. The pontiPy software also computes this chapters equations (Ahn and Verma 2021). Furthermore, the diffeR package in R reads raster maps to compute Quantity, Exchange, and Shift (Pontius Jr and Santacruz 2015).

4.2 Discussion Questions

1. Do the sum of False Alarms across all categories equal the sum of Misses across all categories? Why or why not?
2. Does the difference in the extent equal the sum of False Alarms across all categories plus the sum of Misses across all categories? Why or why not?
3. Under what conditions do the number of observations N_{ij} in each entry influence the marginal sum at the right or at the bottom?
4. What software computes the results for all the equations in this chapter?
5. Transitions that form Exchange between pairs of categories do not involve Quantity of other categories but transitions that form Shift of one category can involve Quantity of other categories, so would it make more conceptual sense for the components' sequence in the figures to be Quantity, Shift, and Exchange rather than Quantity, Exchange, and Shift?

References

Ahn, P., & Verma, P. (2021). pontiPy. https://github.com/verma-priyanka/pontiPy.

Aldwaik, S. Z., & Pontius Jr, R. G. (2012). Intensity analysis to unify measurements of size and stationarity of land changes by interval, category, and transition. *Landscape and Urban Planning, 106*, 103–114. https://doi.org/10.1016/j.landurbplan.2012.02.010.

Aldwaik, S. Z., & Pontius Jr, R. G. (2013). Map errors that could account for deviations from a uniform intensity of land change. *International Journal of Geographical Information Science, 27*, 1717–1739. https://doi.org/10.1080/13658816.2013.787618.

Enaruvbe, G. O., & Pontius Jr, R. G. (2015). Influence of classification errors on intensity analysis of land changes in southern Nigeria. *International Journal of Remote Sensing, 36*, 244–261. https://doi.org/10.1080/01431161.2014.994721.

Pontius Jr, R. G. (2019). Component intensities to relate difference by category with difference overall. *International Journal of Applied Earth Observation and Geoinformation, 77*, 94–99. https://doi.org/10.1016/j.jag.2018.07.024.

Pontius Jr, R. G. (2020). *PontiusMatrix42.xlsx*. http://www.clarku.edu/~rpontius

Pontius Jr, R. G., & Santacruz, A. (2014). Quantity, exchange, and shift components of difference in a square contingency table. *International Journal of Remote Sensing, 35*, 7543–7554. https://doi.org/10.1080/2150704X.2014.969814.

Pontius Jr, R. G., & Santacruz, A. (2015). *diffeR: Metrics of difference for comparing pairs of maps*. https://cran.r-project.org/web/packages/diffeR

Pontius Jr, R. G., Gao, Y., Giner, N., Kohyama, T., Osaki, M., & Hirose, K. (2013). Design and interpretation of intensity analysis illustrated by land change in Central Kalimantan, Indonesia. *Land, 2*, 351–369. https://doi.org/10.3390/land2030351.

Pontius Jr, R. G., Huang, J., Jiang, W., Khallaghi, S., Lin, Y., Liu, J., Quan, B., & Ye, S. (2017). Rules to write mathematics to clarify metrics such as the land use dynamic degrees. *Landscape Ecology, 32*, 2249–2260. https://doi.org/10.1007/s10980-017-0584-x.

Quan, B., Pontius Jr, R. G., & Song, H. (2020). Intensity analysis to communicate land change during three time intervals in two regions of Quanzhou City, China. *GIScience & Remote Sensing, 57*, 21–36. https://doi.org/10.1080/15481603.2019.1658420.

Shafizadeh-Moghadam, H., Minaei, M., Feng, Y., & Pontius Jr, R. G. (2019). GlobeLand30 maps show four times larger gross than net land change from 2000 to 2010 in Asia. *International Journal of Applied Earth Observation and Geoinformation, 78*, 240–248. https://doi.org/10.1016/j.jag.2019.01.003.

Chapter 5
Application to Categorical Error Assessment with Sampling

Abstract The contingency table can derive from a sample of the population, particularly for applications to error assessment. If the contingency table derives from a stratified sample where the strata have various sampling intensities, then it is necessary to convert the sample table to an estimated population table for unbiased assessment. This chapter shows how to perform the conversion, then how to interpret the results when the table's rows are the diagnosed categories and its columns are the reference categories. Relevant software includes the PontiusMatrix42.xlsx spreadsheet available at www.clarku.edu/~rpontius. (Pontius Jr 2020).

Keywords Contingency table · Error assessment · PontiusMatrix · Strata · Sampling

5.1 Text

Consider an application where an algorithm diagnoses a category for each observation. Chapter 1 considered the case where there are exactly two categories: Presence and Absence. This chapter considers the case where there are more than two categories. An example is an algorithm that diagnoses each pixel on a landscape as a category. Error assessment concerns the correspondence between the diagnosed categories and corresponding real categories on the landscape. The profession uses the word reference to refer to the real categories because reality is sometimes difficult to determine. If information concerning the reference categories were easy to collect for the population, then we would not need the diagnosis. Information concerning the reference categories is frequently prohibitively expensive to collect for the population. Therefore, it is common to collect reference information via a sample of the population. The population is the collection of all observations for which we know the diagnosed category, whereas the sample is the subset of observations for which we know both the diagnosed and reference category. Each sample observation has a pair of categories: the diagnosed category and the reference category. A contingency table tallies the number of sampled observations in the table's entries. The table has the diagnosed categories in the rows and the reference categories in the columns, according to the convention in the profession. We want to analyze the

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1_5

data from the sample to gain insight concerning the population. Proper analysis must account for the sampling design, which is the procedure to select the sample. Sampling is an enormous topic that has a wide literature, some of which is specific to land science (Olofsson et al. 2014; Stehman 2020; Stehman and Foody 2019). Many statistics textbooks and university courses fail to teach the concepts of sampling sufficiently for a typical practitioner. This chapter describes fundamental concepts that applied scientists need for applications to error assessment.

Simple random sampling selects observations from the population such that each observation has the same probability of selection. If we use simple random sampling, then the sample contingency table contains unbiased information concerning the relationship between the population's diagnosed categories and the population's reference categories. Thus, if simple random sampling is the design, then the expected values of the summary metrics that derive from the sample contingency table will match to the corresponding values of the summary metrics that would derive from a population contingency table. The metrics from the preceding chapter are examples of such summary metrics. Simple random sampling leads to straightforward analysis of the contingency table, but this does not imply that simple random sampling is a good idea because simple random sampling does not necessarily collect the reference data in the most efficient manner.

A sampling design is more efficient when the design produces information that is more helpful to address the research question, at a given level of effort required to collect the sample. To maximize efficiency, we want each observation to reveal the maximum amount of helpful information per effort required to obtain the reference category for each observation in the sample. Simple random sampling does not necessarily maximize efficiency. For example, simple random sampling will place more samples on larger categories and fewer samples on smaller categories, even when our interest in larger categories is not greater than our interest in smaller categories. If most of the population consists of a large category that we suspect the algorithm diagnoses accurately, then simple random sampling would waste effort in collecting data to confirm something that we already suspect. Simple random sampling causes the sample to have relatively fewer observations for the smaller categories in which we might be more interested. Simple random sampling is not necessarily an efficient sampling design to obtain information that will increase our understanding of how the diagnostic algorithm performs. Furthermore, each observation has the same probability of selection in a simple random sample, but some parts of the population might require more effort than other parts to collect the reference category. For example, some parts of the population might be more difficult to access than other parts. It would be helpful to assign a greater intensity of samples in the parts of the population that give more information and require less effort, while accounting appropriately for the various sizes of the parts. Stratified random sampling is a design that allows us to collect reference data in a manner that generates helpful information more efficiently than simple random sampling.

Stratified random sampling delineates the population into strata. Each stratum is a set where the strata are mutually exclusive and collectively exhaustive. We determine the number of observations to sample from each stratum based on our research goals and labor resources. For example, if the goal is to understand errors, then we would assign more observations to strata where we suspect the diagnosed category is more erroneous. Stratified random sampling then uses simple random sampling in each stratum to select the determined number of observations. The sampling intensity in each stratum is the number of observations in the stratum divided by the size of the stratum. Each observation in a stratum has an equal probability of selection, which is the stratum's sampling intensity. If the sampling intensity varies among strata, then the probability of an observation's selection varies among strata. Therefore, analysis of data from a stratified sampling design requires an additional step that data from a simple random sampling design does not require. If some strata have a different sampling intensity than other strata, then we must convert the sample table into an estimated population table. If we fail to convert the sample table into an estimated population table, then the resulting summary metrics might be biased. I have seen numerous examples where scientists produced biased results because scientists failed to convert their sample tables into estimated population tables. Software packages that lack a tool for conversion lead scientists into this common blunder. The equations below give a straightforward method to convert the sample table into the estimated population table, which generates unbiased summary metrics.

Table 5.1 gives the mathematical notation. We separate the population into strata, where B denotes the number of strata and b denotes a particular stratum. N_b denotes the size of the population in stratum b. We assign more observations to the strata that we suspect are more informative per effort required to collect the reference data. Thus, the strata that are more important to our research question will be the strata that have a greater sampling intensity. Equation 5.1 defines the sampling intensity for each stratum as the number of observations in the stratum divided by the size of the stratum. We use simple random sampling in each stratum to select the stratum's assigned number of observations. We record each observation as category i for variable **X** and category j for variable **Y** then compile a sample contingency table where **X** is the diagnosed category and **Y** is the reference category.

Table 5.1 Notation to estimate population contingency table for stratified sampling

Notation	Meaning
b	Index for a stratum
B	Number of strata
i	Index for a category in table's rows where $i = 1, 2, \ldots J$
j	Index for a category in table's columns where $j = 1, 2, \ldots J$
J	Number of categories
n_{bij}	Number of sampled observations in stratum b that are row i and column j
n_{ij}	Number of sampled observations in stratum i that are row i and column j
N_b	Size of population in stratum b
N_i	Size of population in stratum i when the strata are the row categories
$\hat{N}_{ij}$	Entry in estimated population contingency table for row i and column j
S_b	Sampling intensity in stratum b

Equation 5.2 converts the sample contingency table into an estimated population table. Equation 5.2 gives each entry in the estimated population table, where each entry is the estimated size of the population that is category i for variable $\mathbf{X}$ and category j for variable $\mathbf{Y}$. Equation 5.2 accounts for the possibility that the sampling intensities might vary among strata because the equation is a weighted sum where weight for stratum b is the inverse of the sampling intensity for stratum b. If the strata are the row categories, then Eq. 5.2 simplifies to Eq. 5.3, where N_i denotes size of stratum i. Equation 5.3 is analogous to Eq. 1 in Pontius Jr and Millones (2011).

$$S_b = \frac{\sum_{i=1}^{J}\sum_{j=1}^{J} n_{bij}}{N_b} \tag{5.1}$$

$$\hat{N}_{ij} = \sum_{b=1}^{B}\left(\frac{n_{bij}\, N_b}{\sum_{i=1}^{J}\sum_{j=1}^{J} n_{bij}}\right) = \sum_{b=1}^{B}\left(\frac{n_{bij}}{S_b}\right) \tag{5.2}$$

$$\hat{N}_{ij} = \frac{n_{ij}\, N_i}{\sum_{j=1}^{J} n_{ij}} \text{ when strata are the row categories} \tag{5.3}$$

Figure 5.1 presents an example where three strata are the three row categories, thus Eq. 5.3 applies. The population size is 1000. Row categories 1 and 2 each account for 240 of the population's 1000 while row category 3 accounts for 520. Suppose we have sufficient labor resources to collect reference information for 74 observations. We must allocate those 74 samples among the three strata. Figure 5.1 shows an example where the stratified sample has 24 observations in each of strata 1 and 2, thus those two strata have a sampling intensity of 10%. Stratum 3 has 26 observations, thus stratum 3 has a sampling intensity of 5%. There are many reasons why we might allocate the number of samples in each stratum with unequal sampling intensities among the strata. It could be that strata 1 and 2 are more informative than stratum 3 for our goals or maybe stratum 3 requires more effort than strata 1 and 2 to collect the reference data. The sample contingency table compiles the sample data. Equation 5.3 converts the sample contingency table into the estimated population table.

The estimated population table generates unbiased estimates of summary metrics because the estimated population table accounts for the unequal sampling intensities among the strata. The sums at the right of the tables show the sizes of the various strata, while the sums at the bottom of the estimated population table show that the estimated size of category 3 is eight times larger than categories 1 and 2. The central entries in the table describe the association between the diagnosis and the reference. Each off-diagonal entry is a False Alarm for the diagnosed row category and a Miss for the reference column category. For our example, the entries in each column indicate that the algorithm diagnosed the correct category with 60% accuracy when the algorithm encountered the column's category. When the algorithm missed the correct category in a column, the algorithm diagnosed equal sizes of

Raw Sample

		Y = Reference				
		j=1	*j*=2	*j*=3	**Sum**	**Stratum**
X=Diagnosis	*i*=1	6	2	16	**24**	**240**
	i=2	2	6	16	**24**	**240**
	i=3	1	1	24	**26**	**520**

Expected Population

		Y=Reference				
		j=1	*j*=2	*j*=3	**Sum**	**False Alarms**
X=Diagnosis	*i*=1	60	20	160	**240**	**180**
	i=2	20	60	160	**240**	**180**
	i=3	20	20	480	**520**	**40**
	Sum	**100**	**100**	**800**	**1000**	**400**
	Misses	**40**	**40**	**320**	**400**	

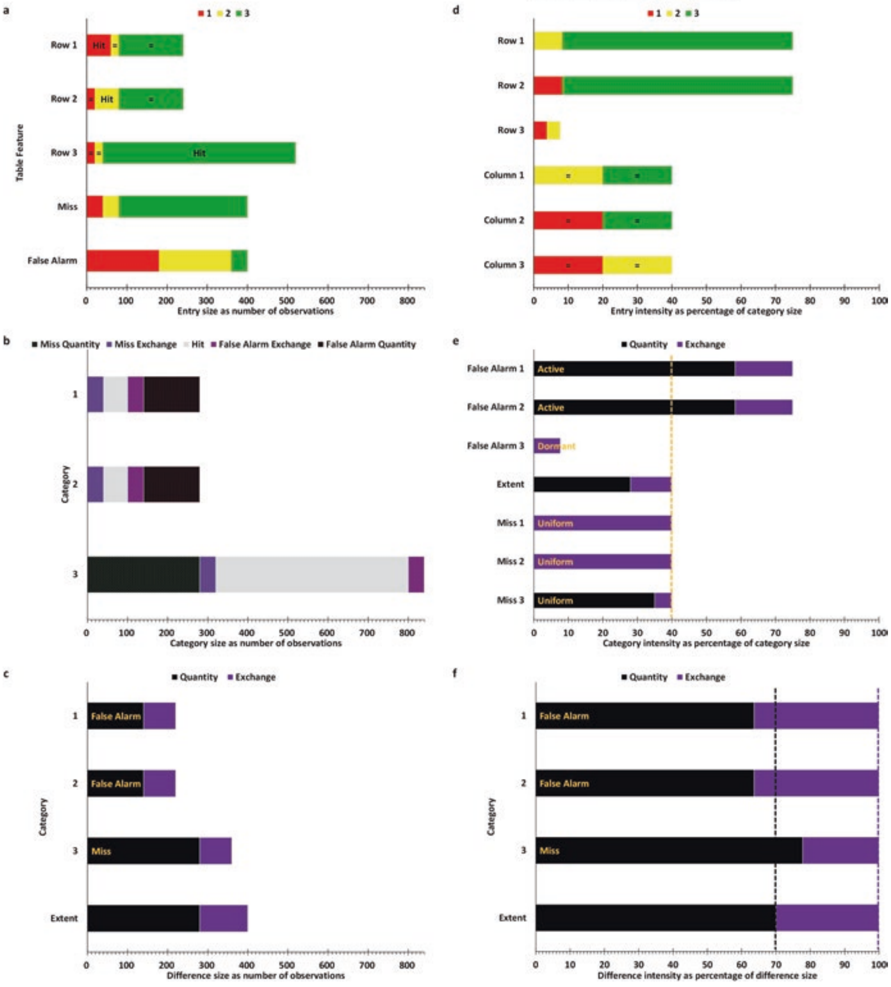


Fig. 5.1 Application to stratified sampling

False Alarms in the rows of the incorrect categories. Category 3 is eight times larger than the other categories in the reference information, thus the Misses of category 3 are eight times larger than the Misses of the other categories.

Figure 5.1a–c shows the estimated population table's sizes. Figure 5.1a gives the sizes of the entries in the population table. The labels on the segments in Fig. 5.1a indicate the relative sizes of the off-diagonal entries in each column. For example, the two equals signs in the red segments indicate that the size of both False Alarms equal each other in column 1. This indicates that when the diagnostic algorithm encounters reference category 1, the algorithm diagnoses categories 2 and 3 with equal size, which is 20. Similarly, the equals signs on the yellow segments indicate that when the algorithm encounters reference category 2, the algorithm generates False Alarms for categories 1 and 3 with equal size. The equals signs on the green segments indicate that when the algorithm misses category 3, the algorithm generates False Alarms for categories 1 and 2 with equal size, which is 160. The sum of Misses must equal the sum of False Alarms because each error is a Miss of one category and a False Alarm of a different category. Figure 5.1b shows a rectangular Venn diagram for each category, where the set on the left of each Venn diagram is the reference and the set on the right is the diagnosis. Hits are the intersection of reference and diagnosis. The Miss Exchange component must be the same size as the False Alarm Exchange component for each category. The Shift components are zero in this example. Categories 1 and 2 have a positive False Alarm Quantity component, which indicates that the diagnosis shows more of categories 1 and 2 than exist in the reference. Category 3 has a positive Miss Quantity component, which indicates that the diagnosis shows less of category 3 than exists in the reference. Figure 5.1c shows the sizes of the difference components. The extent bar shows the diagnosis differs from the reference for an estimated 400 observations, and the Quantity component accounts for most of that difference. Category 1 has 220 observations of difference, which is the sum of Misses and False Alarms for category 1. False Alarms are larger than Misses for category 1, thus category 1 has a Quantity component with the label False Alarm. The results for category 1 are identical to the results for category 2. Misses are larger than False Alarms for category 3, thus Miss is the label on the Quantity component for category 3.

Figure 5.1d–f shows the estimated population table's intensities. The bars titled as row in Fig. 5.1d reveal each category's False Alarm intensity while the bars titled as column reveal each category's Miss intensity. Row 1 indicates that 75% of the observations diagnosed as category 1 are False Alarms, most of which are Misses of category 3. Row 2 shows the same pattern. Row 3 indicates that about 8% of the diagnoses of category 3 are False Alarms, split equally between Misses of categories 1 and 2. The bars titled columns in Fig. 5.1d indicate all categories have Miss intensity of 40%, where each category's Misses are distributed equally as False Alarms of the other two categories. The labels indicate how the False Alarms are distributed in a row with respect the sizes of the reference categories. The labels derive from Eqs. 4.19 and 4.39. The distribution of the False Alarms in row 1 are proportional to the sizes of categories 2 and 3 in the reference, so both of the red segments have equals signs in Fig. 5.1d. The False Alarms in row 2 are proportional to the sizes of categories 1 and 3 in the reference, so both of the yellow segments have equals signs. The False Alarms in row 3 are proportional to the sizes of categories 1 and 2 in the reference information, so both of the green segments have equals signs. Figure 5.1e shows the categories' intensities compared to the error intensity overall in the extent, which the dashed line shows as 40%. The False Alarm intensities for categories 1 and 2 are greater than 40%, so categories 1 and 2 have the label

Active. The False Alarm intensity for category is less than 40%, so category 3 has the label Dormant. The Miss intensities for all categories are 40%, so all categories have the label Uniform for their Miss intensities. Figure 5.1e show also the portion of each intensity that derives from the components of Quantity and Exchange. Figure 5.1f shows the components as a percentage of the difference. The dashed lines show that 70% of the extent's difference derives from the Quantity component and 30% derives from the Exchange component. Approximately 64% of the difference in categories 1 and 2 derive from their Quantity component, whereas 80% of the difference in category 3 derives from its Quantity component. Thus categories 1 and 2 derive their difference from their Quantity components less intensively than the extent derives its difference from its Quantity component. Category 3 derives its difference from its Quantity components more intensively than the extent derives the extent's difference from the extent's Quantity component.

The PontiusMatrix42.xlsx spreadsheet computed the results in Fig. 5.1 by reading the sample table and the size of the strata (Pontius Jr 2020). The spreadsheet automatically converts a sample table into the corresponding population table, and then generates graphs from the population table. Figure 5.1 offers numerous measurements, where each measurement addresses a particular research question. We must interpret the metrics in the context questions that apply to error assessment. The first thing that might interest us is the overall percentage of error. The extent bar in Fig. 5.1e shows that the overall error intensity is 40% of the population. The overall error from the sample table is 52%, which illustrates the importance of converting the sample table into the estimated population table. If we were to have computed the summary metrics from the sample table, then we would have obtained a biased estimate of the overall percentage error. The sampling intensities in strata 1 and 2 are double the sampling intensity in stratum 3, thus strata 1 and 2 would be overrepresented in summary metrics that derive directly from the sample table. Strata 1 and 2 have a larger False Alarm intensity than stratum 3, thus the overall percentage error from the sample table is larger than from the population.

The producer of the diagnostic algorithm would be especially interested in the types of errors the algorithm makes when the algorithm encounters each reference category in column *j*. The column bars in Fig. 5.1d give this information. The algorithm produces a Miss intensity of 40% for each category in column *j*. One minus the Miss intensity for category *j* is known as the producer's accuracy for category *j* because this metric would interest the producer of the diagnostic algorithm. Figure 5.1d shows further the False Alarm category in row *i* for the Misses of *j*.

The user of the diagnosis would be especially interested in the types of errors the algorithm makes when the algorithm diagnoses each category in row *i*. The row bars in Fig. 5.1d give this information. The algorithm produces a smaller False Alarm intensity for category 3 than for the other two categories. One minus the False Alarm intensity for category *i* is known as the user's accuracy for category *i* because this metric would interest the user of the diagnosis.

Figure 5.1d, e show that the Miss intensities for the reference categories are equal to each other while the False Alarm intensities for the diagnosis categories are not equal to each other. This illustrates how the sizes of the reference categories influence the sizes of the False Alarms in each row *i*, while the entries in the table do not influence the size of the reference categories. The diagnostic algorithm treats each reference category in the same manner in each column but the population table's entry in row 1 column 3 is eight times larger than the entry in row 1 column 2 because the size of category 3 in the reference is eight times the size of category 2 in the reference. The sizes of the False Alarms in each row are directly proportional to the sizes of the reference categories. Therefore, equals signs appear in the segments of the column bars in Fig. 5.1d because the result from Eq. 4.39 for row *i* is identical to the results from Eq. 4.19 applied to each row *i*.

The diagnostic algorithm responds identically to each reference category, thus the Miss intensities are identical for each category *j*. Therefore, the larger reference categories have more Misses than the smaller reference categories. The sizes of the reference categories influence the sizes of the False Alarms in each row *i* thus Fig. 5.1e shows that the False Alarm intensities are not identical for each row *i*. For example, the False Alarm intensity of category 3 is dormant while the False Alarm intensities are active for categories 1 and 2 because the reference category 3 is larger than the other reference categories. If we were to apply the same diagnostic algorithm to a population that has a different distribution of sizes of reference categories, then we would expect different False Alarm intensities because the sizes of the reference categories influences the sizes in the diagnosis. However, we would expect the same Miss intensities because the diagnosis does not influence the reference. This is why the producer's of the algorithm would be more interested in the Miss intensities than in the False Alarm intensities. The Miss intensities reflect the ability of the algorithm to distinguish the various reference categories, regardless of the sizes of the reference categories in any particular population. The sizes of the reference categories influence the False Alarm intensities, which give results that are specific to the particular population. The False Alarm intensities would be interesting to a user who is interested in the diagnosis for the particular population.

The labels on the Quantity components in Fig. 5.1 c, f illustrate another way that the reference sizes in the population influence the results. The Quantity component for categories 1 and 2 have a label of False Alarm because False Alarms are greater than Misses for those categories, which indicates the diagnosis overestimates the sizes of those categories. The Quantity component for category 3 has a label of Miss, which indicates that the diagnosis underestimates the sizes of category 3. The overestimation of categories 1 and 2 and the underestimation of category 3 is because of the unbalanced sizes of the reference categories, not because of the diagnostic algorithm. If the algorithm were applied to a population with balanced sizes of the three categories, then the diagnosis would not underestimate or overestimate any category, because the algorithm treats each category in the same manner concerning how the algorithm misses each category.

This chapter concerns sampling, therefore inferential statistics apply. This chapter does not include equations to compute confidence intervals and to perform hypothesis tests for the estimated metrics. Readers can find such equations for the most popular metrics in the literature (Olofsson et al. 2014; Stehman 2020; Stehman and Foody 2019). Such equations for some of this book's novel metrics do not yet exist.

5.2 Discussion Questions

1. Under what conditions is it necessary to convert the sample contingency table to an estimated population contingency table?
2. What problems will you encounter if you do not convert the sample contingency table to an estimated population contingency table when using a stratified sampling design with inconsistent sampling intensity in the various strata?
3. What would motivate a scientist to use a simple random sample?
4. Under what conditions is a simple random sample less efficient than a stratified random sample?
5. Why would a scientist apply unequal sampling intensities to the various strata?
6. Is it necessary that the strata correspond to the categories in the rows of the contingency table?
7. Why would the producer of an algorithm be more interested in the Miss intensities than the False Alarm intensities?
8. If the application of a diagnostic algorithm underestimates or overestimates the size of a category, then should the algorithm's producer modify the algorithm to fix the Quantity error?
9. How can you use PontiusMatrix42.xlsx to convert a stratified sample table into an estimated population table?

References

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment, 148*, 42–57. https://doi.org/10.1016/j.rse.2014.02.015.

Pontius Jr, R. G. (2020). *PontiusMatrix42.xlsx*. http://www.clarku.edu/~rpontius

Pontius Jr, R. G., & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing, 32*, 4407–4429. https://doi.org/10.1080/01431161.2011.552923.

Stehman, S. V. (2020). Ground verification and accuracy assessment. In J. P. Wilson (Ed.), *The geographic information science & technology body of knowledge*. Washington, DC: Association of American Geographers.

Stehman, S. V., & Foody, G. M. (2019). Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment, 231*, 111199. https://doi.org/10.1016/j.rse.2019.05.018.

Chapter 6
Multiple Spatial Resolutions for Categorical Variables

Abstract Analysis at multiple spatial resolutions allows insight concerning the spatial relationship between False Alarms and Misses for a category. A coarsening algorithm converts fine-resolution observations into blocks that have a coarser spatial resolution, which can cause each block to have membership to more than one category. This chapter shows how to construct a square contingency table when a block can have membership to more than one category. Then the concepts of Chap. 4 analyze the contingency table to compute results at each resolution. The sum of Exchange and Shift shrinks as the spatial resolution grows coarser. The spatial resolutions at which these components shrink gives insight to the spatial allocation of a category's False Alarms and Misses. Relevant software includes the CROSSTAB module in TerrSet available at https://clarklabs.org and the diffeR package available at https://cran.r-project.org/web/packages/diffeR/index.html.

Keywords Allocation difference · Contingency table · Multiple-resolution · PontiusMatrix

6.1 Text

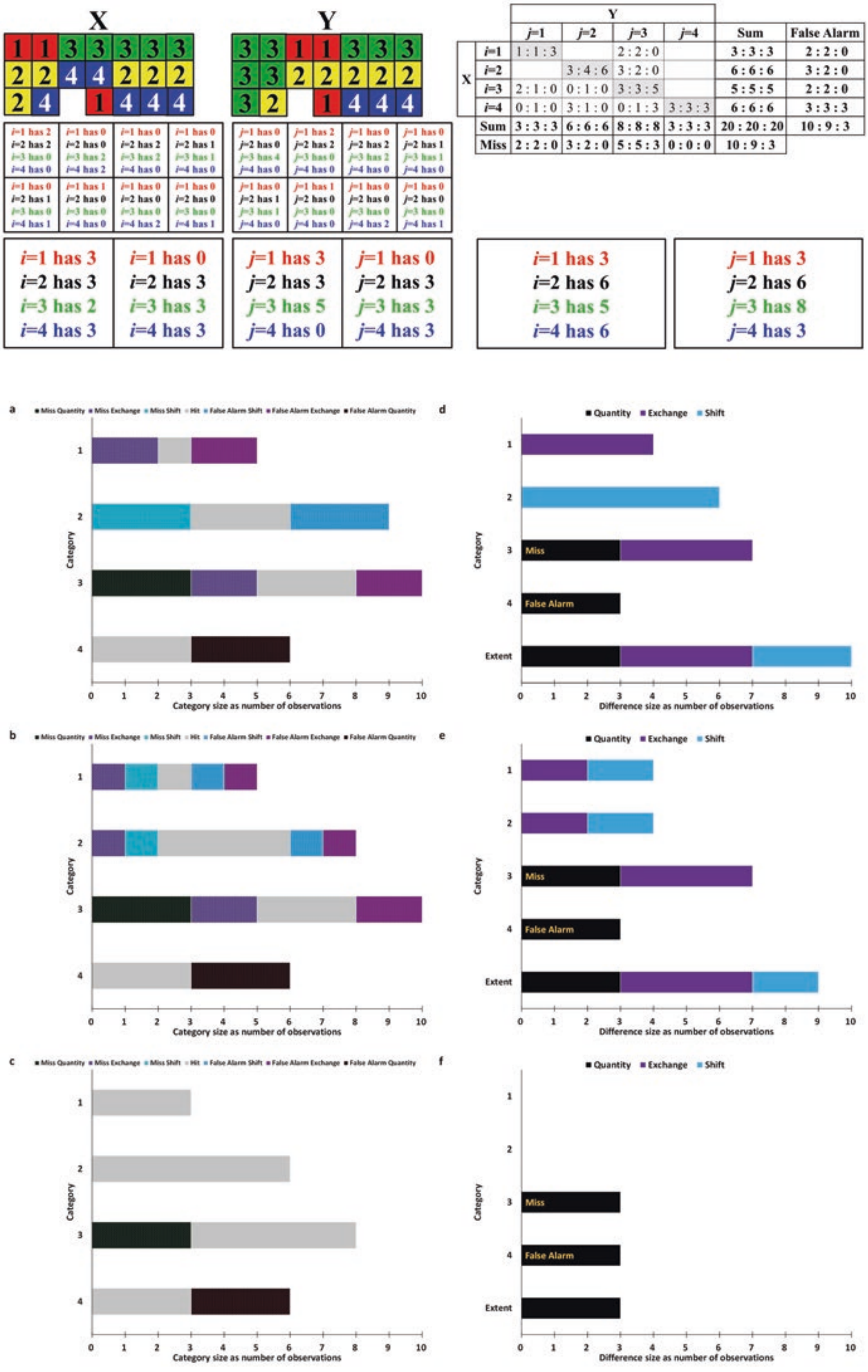
The example in Chap. 4 shows how to compare maps comprised of pixels. The pixel-by-pixel approach of Chap. 4 compares each pixel's **X** category with the pixel's corresponding **Y** category. If a category has False Alarms in some pixels and Misses in other pixels, then the category demonstrates Allocation difference, which is the sum of the Exchange and Shift components. Near Allocation difference exists where a category's False Alarm is near the category's Miss. Far Allocation difference exists where a category's False Alarm is far from the category's Miss. It is helpful to distinguish between near and far Allocation differences of various applications (Pontius Jr et al. 2007, 2008). Pixel-by-pixel analysis of Chap. 4 lacks the ability to distinguish between near and far Allocation differences. Multiple-resolution analysis can distinguish between near and far Allocation differences. This chapter describes how to perform and to interpret multiple-resolution analysis.

The data from Chap. 4 has 20 observations, where each observation has complete membership to exactly one category. Those data exist in space such that the distance

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1_6

between some pairs of observations differs from the distance between other pairs of observations. The methods of Chap. 4 ignore the spatial proximity among the observations. This chapter accounts for spatial proximity by analyzing the data at multiple coarser spatial resolutions. The top of Fig. 6.1 shows the fine-resolution observations. Below the fine-resolution data are the same data at coarser resolutions. The coarsening algorithm begins at the upper left corner to merge each square cluster of fine-resolution observations into a coarser block. Figure 6.1 shows four resolutions: 1, 2, 4, and 8. Resolution 1 is the resolution of the fine-resolution data. The length of the side of each block at resolution 2 is two times the length of each observation at resolution 1. The membership to a category in each block is the sum of memberships to the category of the observations that constitute the block. For the block in the upper left at resolution 2, **X** has membership of two to category 1 and membership of two to category 2, while the corresponding **Y** block has membership of four to category 3. The length of the side of each block at resolution 4 is four times the length of the observation at resolution 1. At resolution 4, **X** differs from **Y** in the left block, while **X** is identical to **Y** in the right block. At resolution 8, the entire extent is in one block. We can use multiple-resolution analysis to distinguish near from far Allocation difference where near means inside the blocks, where each resolution has a specific size of the blocks. The coarsening can cause each block to have membership to more than one category, thus construction of the contingency table requires deep thought, which this chapter gives. Table 6.1 gives the mathematical notation to construct the contingency table at each resolution. The upper right of Fig. 6.1 shows the contingency table where colons separate each entry for resolutions 1, 2, and 4.

Equations 6.1, 6.2, 6.3, 6.4, 6.5 and 6.6 show how to construct the contingency table for cases when each block can have membership to more than one category. The method computes a table for each block, and then sums the tables over all blocks. The approach is similar to how Eq. 5.2 compiles a table from strata. The strata of Chap. 5 are the blocks of this Chap. 6. Each stratum in Chap. 5 generates has its own contingency table, and each block of this Chap. 6 has its own contingency table. However, the contingency table for each stratum includes the Allocation difference in each stratum, while the contingency table for each block ignores the Allocation difference in each block. Equation 6.1 gives a constraint that the sum of the memberships to all categories in each block for **X** must equal the sum of the memberships to all categories in the corresponding block for **Y**. Equation 6.2 defines the size of a block's Hit for each category, which appears on a diagonal entry of the block's table. The MINIMUM operator assures that the diagonal entry for a category in the table is never larger than the size of the category in **X** or in **Y**. If the size of a category in **X** matches the size of the same category in **Y**, then the MINIMUM operator assigns that size as the Hit for that category. The MINIMUM operator allocates the memberships of **X** and **Y** as much as possible to the diagonal entries in the contingency table. Equation 6.2 computes the Hit for each category, then Eq. 6.3 computes the False Alarm while Eq. 6.4 computes the Miss for each category. Equation 6.3 defines False Alarms so that the marginal sums at the right of the table equal the sum of memberships to the categories in **X**. Equation 6.4 defines Misses



X						
1	1	3	3	3	3	3
2	2	4	4	2	2	2
2	4		1	4	4	4

Y						
3	3	1	1	3	3	3
3	3	2	2	2	2	2
3	2		1	4	4	4

i=1 has 2
i=2 has 2
i=3 has 0
i=4 has 0
	i=1 has 0
	i=2 has 0
	i=3 has 2
	i=4 has 2
		i=1 has 0
		i=2 has 2
		i=3 has 2
		i=4 has 0
			i=1 has 0
			i=2 has 1
			i=3 has 1
			i=4 has 0

i=1 has 0
i=2 has 1
i=3 has 0
i=4 has 1
	i=1 has 1
	i=2 has 0
	i=3 has 0
	i=4 has 0
		i=1 has 0
		i=2 has 0
		i=3 has 0
		i=4 has 2
			i=1 has 0
			i=2 has 0
			i=3 has 0
			i=4 has 1

j=1 has 0
j=2 has 0
j=3 has 4
j=4 has 0
	j=1 has 2
	j=2 has 2
	j=3 has 0
	j=4 has 0
		j=1 has 0
		j=2 has 2
		j=3 has 2
		j=4 has 0
			j=1 has 0
			j=2 has 1
			j=3 has 1
			j=4 has 0

j=1 has 0
j=2 has 1
j=3 has 1
j=4 has 0
	j=1 has 1
	j=2 has 0
	j=3 has 0
	j=4 has 0
		j=1 has 0
		j=2 has 0
		j=3 has 0
		j=4 has 2
			j=1 has 0
			j=2 has 0
			j=3 has 0
			j=4 has 1

i=1 has 3
i=2 has 3
i=3 has 2
i=4 has 3
	i=1 has 0
	i=2 has 3
	i=3 has 3
	i=4 has 3

j=1 has 3
j=2 has 3
j=3 has 5
j=4 has 0
	j=1 has 0
	j=2 has 3
	j=3 has 3
	j=4 has 3

		Y					
		j=1	j=2	j=3	j=4	Sum	False Alarm
X	i=1	1 : 1 : 3		2 : 2 : 0		3 : 3 : 3	2 : 2 : 0
	i=2		3 : 4 : 6	3 : 2 : 0		6 : 6 : 6	3 : 2 : 0
	i=3	2 : 1 : 0	0 : 1 : 0	3 : 3 : 5		5 : 5 : 5	2 : 2 : 0
	i=4	0 : 1 : 0	3 : 1 : 0	0 : 1 : 3	3 : 3 : 3	6 : 6 : 6	3 : 3 : 3
	Sum	3 : 3 : 3	6 : 6 : 6	8 : 8 : 8	3 : 3 : 3	20 : 20 : 20	10 : 9 : 3
	Miss	2 : 2 : 0	3 : 2 : 0	5 : 5 : 3	0 : 0 : 0	10 : 9 : 3	

i=1 has 3
i=2 has 6
i=3 has 5
i=4 has 6
	j=1 has 3
	j=2 has 6
	j=3 has 8
	j=4 has 3

a
■ Miss Quantity ■ Miss Exchange ■ Miss Shift ■ Hit ■ False Alarm Shift ■ False Alarm Exchange ■ False Alarm Quantity
Category
1
2
3
4
0 1 2 3 4 5 6 7 8 9 10
Category size as number of observations
d
■ Quantity ■ Exchange ■ Shift
Category
1
2
3
4
Extent
Miss
False Alarm
0 1 2 3 4 5 6 7 8 9 10
Difference size as number of observations
b
■ Miss Quantity ■ Miss Exchange ■ Miss Shift ■ Hit ■ False Alarm Shift ■ False Alarm Exchange ■ False Alarm Quantity
Category
1
2
3
4
0 1 2 3 4 5 6 7 8 9 10
Category size as number of observations
e
■ Quantity ■ Exchange ■ Shift
Category
1
2
3
4
Extent
Miss
False Alarm
0 1 2 3 4 5 6 7 8 9 10
Difference size as number of observations
c
■ Miss Quantity ■ Miss Exchange ■ Miss Shift ■ Hit ■ False Alarm Shift ■ False Alarm Exchange ■ False Alarm Quantity
Category
1
2
3
4
0 1 2 3 4 5 6 7 8 9 10
Category size as number of observations
f
■ Quantity ■ Exchange ■ Shift
Category
1
2
3
4
Extent
Miss
False Alarm
0 1 2 3 4 5 6 7 8 9 10
Difference size as number of observations

Fig. 6.1 Example to show blocks for multiple resolutions

Table 6.1 Notation to compute the table when each block can contain more than one category

Notation	Meaning
b	Index for a block where $b = 1, 2, \ldots B$
B	Number of blocks
F_{bi}	Size of False Alarm for block b in table's row i
H_{bi}	Size of Hit for block b in table's row i
H_{bj}	Size of Hit for block b in table's column j
i	Index for a category of **X** in table's row where $i = 1, 2, \ldots J$
j	Index for a category of **Y** the table's column where $j = 1, 2, \ldots J$
J	Number of categories
M_{bj}	Size of Miss for block b in table's column j
S_{bii}	Size of entry for block b in table's row i and column i
S_{bij}	Size of entry for block b in table's row i and column j
S_{bjj}	Size of entry for block b in table's row j and column j
N_{ij}	Size of entry in row i and column j of overall table at blocks' resolution
X_{bi}	Size of membership in **X** of block b to category i
Y_{bj}	Size of membership in **Y** of block b to category j

so that the marginal sums at the bottom of the table equal the sum of memberships in **Y**. The definition of Hit causes the False Alarm and/or the Miss of each category to be zero in each block. Equation 6.5 allocates the False Alarm of i to each off-diagonal entry in proportion to the relative sizes for Misses of j. Equation 6.5 is equivalent to allocating the Miss of j to each off-diagonal entry in proportion to the relative sizes for False Alarms for i. Multiplication in the numerator of Eq. 6.5 performs an allocation that treats each category in an identical mathematical manner. Each block generates a contingency table where Eq. 6.2 gives the table's diagonal entries and Eq. 6.5 gives the table's off-diagonal entries according to rules known as the Composite operator (Kuzera and Pontius Jr 2008; Pontius Jr and Cheuk 2006; Pontius Jr and Connors 2009). Equation 6.6 sums each entry over all blocks to give the overall table at the resolution of the blocks. The concepts of Chap. 4 can then analyze the overall table.

$$\sum_{i=1}^{J} X_{bi} = \sum_{j=1}^{J} Y_{bj} \tag{6.1}$$

$$S_{bii} = S_{bjj} = H_{bi} = H_{bj} = \text{MINIMUM}\left(X_{bi}, Y_{bj}\right) \text{when}\, i = j \tag{6.2}$$

$$F_{bi} = X_{bi} - H_{bi} \tag{6.3}$$

$$M_{bj} = Y_{bj} - H_{bj} \tag{6.4}$$

$$S_{bij} = \frac{F_{bi}\, M_{bj}}{\sum_{j=1}^{J} M_{bj}} = \frac{F_{bi}\, M_{bj}}{\sum_{i=1}^{J} F_{bi}}\, when\, i \neq j \tag{6.5}$$

$$N_{ij} = \sum_{b=1}^{B} S_{bij} \tag{6.6}$$

The conversion from the fine resolution to a coarser resolution might affect the entries in the overall table, depending on the spatial allocation of the fine-resolution observations. The reader should use paper and pencil to compute the table for each block to learn the relationship between spatial allocation and multiple resolutions. The following three paragraphs describe each of three blocks at resolution 2, which illustrate how Eqs. 6.1, 6.2, 6.3, 6.4, 6.5 and 6.6 work.

The block in the top left at resolution 2 derives from four observations at resolution 1. The four observations contribute values of two in row 1 column 3 and two in row 2 column 3 to the table at resolution 1. The top left block at resolution 2 has membership to more than one category for **X**, while the block has membership entirely to category 3 for **Y**. Equation 6.2 computes values of zero for each diagonal entry for the block. Equation 6.5 computes two in row 1 column 3 and two in row 2 column 3, while Eq. 6.5 computes zero in the other off-diagonal entries at resolution 2. Thus, the coarsening from resolution 1 to resolution 2 for the top left block has no effect on the overall table's entries.

In contrast, coarsening from resolution 1 to 2 affects the table's entries in the block second from left in the top of the extent. That block derives from four fine-resolution observations that contribute two in both row 3 column 1 and row 4 column 2. The coarsening causes the block to have membership to more than one category for both **X** and **Y** at resolution 2. Equation 6.2 computes zero for all diagonal entries for that block because **X** and **Y** do not have any categories in common. Equation 6.5 computes one for each of the four entries in row 3 column 1, row 3 column 2, row 4 column 1, and row 4 column 2. Coarsening causes dispersion among the off-diagonal values in the table. The dispersion in the table reflects the dispersion of categories in the block. The coarsening causes the entry in row 3 column 2 to grow, which reflects that category 3 in **X** is near category 2 in **Y** at the fine resolution. Similarly, coarsening causes the entry in row 4 column 1 to grow, which reflects that category 4 in **X** is near category 1 in **Y** at the fine resolution. The word near means in a block. This illustrates how the coarsening can cause the values in some off-diagonal entries to migrate to other off-diagonal entries.

The block in the bottom left at resolution 2 derives from two fine-resolution observations. This illustrates that it is not necessary for the sum of memberships in one block to equal the sum of memberships in other blocks. The two fine-resolution observations have a False Alarm for category 2 and a Miss for category 2. The merger of those two fine-resolution observations produces one block that has membership of one to category 2 for both **X** and **Y**. Equation 6.2 computes a value of one on the diagonal for category 2, meaning a Hit for category 2. If a fine-resolution False Alarm for a category is near a Miss for the category, then the False Alarm and the Miss form a Hit for the category, where near means in a block. For this situation, coarsening causes an off-diagonal value to migrate to the diagonal of the table, thus difference shrinks and Hits grow. Equation 6.5 computes one in row 4 column 3 for the block at resolution 2. The coarsening causes growth of the entry in row 4 column 3, because the fine-resolution False Alarm of category 4 is near the fine resolution Miss of category 3.

Coarsening to resolution 2 does not affect the right half of the spatial extent, because the right half consists entirely of Hits at the fine resolution. Equation 6.2 assigns all of the memberships to the diagonal entries for the blocks that form the right half of the extent. Consequently, Eq. 6.5 has no remaining membership to allocate to the table's off-diagonal entries.

The conversion to a coarser resolution does not influence the size of each category in **X** or in **Y**, thus does not influence the sums at the right or the bottom of the overall contingency table. Coarsening maintains the sizes of the categories but reduces the precision concerning the spatial allocation of the categories. Thus coarsening maintains the table's marginal sums but can influence the allocation of the entries in the contingency table. Comparison of tables across various resolutions reveals information concerning exclusively the spatial allocation of the observations at the fine resolution.

Resolution 4 consists of two blocks: a left block and a right block. Variables **X** and **Y** differ in the left block, but **X** and **Y** are identical in the right block. Equation 6.2 computes values of three, three, two, and zero for the respective diagonal entries of categories 1, 2, 3, and 4 for the left block. Equation 6.5 computes a single positive off-diagonal value of three in row 4 column 3 for the left block. For the right block, Eq. 6.2 computes values of zero, three, three, and three for the respective diagonal entries of categories 1, 2, 3, and 4, then Eq. 6.5 computes zero for all off-diagonal entries. Comparison of the table at resolution 1 with the table at resolution 4 show how off-diagonal values at resolution 1 migrate to the diagonal at resolution 4, thus difference shrinks and Hits grow as resolution becomes coarser. Resolution 8 contains the entire extent in one block and produces a table identical to the overall table at resolution 4.

Equation 6.6 sums the blocks at each resolution. The upper right part of Fig. 6.1 gives the overall contingency table in which a colon separates resolutions 1, 2, and 4. Then the equations from Chap. 4 compute summary metrics at each resolution. The bottom of Fig. 6.1 shows the difference components at each resolution for each category and overall. The total difference for each category and overall remains the same or shrinks as resolution becomes coarser. If a finer resolution is nested in a coarser resolution as Fig. 6.1 shows, then it is impossible for the coarsening to cause the total difference to grow. Figure 6.1 shows the coarsening causes the overall difference to shrink from 10 to 9 to 3. Category 2 is responsible for the shrinking of overall difference from resolution 1 to resolution 2 because category 2 has a False Alarm near a Miss at the fine resolution, where near means in a block at resolution 2. Coarsening never influences the Quantity components, while the sum of Exchange and Shift components shrinks or remains constant. The following paragraphs describe how coarsening can cause the Exchange or Shift components to shrink or grow. In the following paragraphs, a False Alarm of *i* refers to category *i* in **X** while a Miss of *j* refers to category *j* in **Y**.

Consider the coarsening from resolution 1 to 2. Category 1 has a difference of four, consisting of Exchange with category 3 at resolution 1. At resolution 2, category 1 still has a difference of four but is split between Exchange and Shift. At resolution 1, two Misses of category 1 are co-located with False Alarms of 3. The coarsening from resolution 1 to 2 causes the two Misses of category 1 to become associated with False Alarms of both 3 and 4. Consequently, Exchange shrinks and Shift grows for category 1 as resolution becomes coarser. Now consider category 2, which has a difference of four, consisting of Shift with categories 3 and 4 at resolution 1. Three Misses of 2 are co-located with False Alarms of 4 at resolution 1, but two Misses of 2 are near False Alarms of 3. The coarsening to resolution 2 causes the Misses of 2 to have stronger association with the False Alarms of 3 and weaker association with False Alarms of 4, thus category 2 experiences growth of Exchange and shrinkage of Shift.

The coarsening from resolution 1 to 2 does not influence the components for category 3, but the coarsening modifies how category 3 exchanges with particular categories. False Alarms of 3 are co-located with Misses of 1 at resolution 1. At resolution 2, the False Alarms of 3 become associated with Misses of 1 and 2. The tables show that category 3 exchanges with only category 1 at resolution 1. At resolution 2, category 3 exchanges equally with categories 1 and 2.

Category 4 has False Alarms but no Misses at resolution 1, thus all of the difference for category 4 is the Quantity component, which coarsening does not influence. However, coarsening influences how the False Alarms in row 4 are distributed among the table's columns at each resolution. The False Alarms of 4 migrate from the Miss of 2 at the resolution 1 to the Miss of 3 at resolution 4.

The overall table at resolution 4 shows that no single category has both False Alarms and Misses, thus the Exchange and Shift components are zero. Comparison among the tables reveals information concerning the spatial allocation of the differences, specifically that all of the Exchange and Shift occurs in the left side of the spatial extent, which Fig. 6.1 shows. Consequently, results at resolution 8 are identical to the results at resolution 4.

The diffeR package in R reads raster maps to compute Quantity, Exchange, and Shift at multiple resolutions (Pontius Jr and Santacruz 2015). The package is free at https://cran.r-project.org/web/packages/diffeR.

6.2 Discussion Questions

1. What information can multiple-resolution analysis reveal that pixel-by-pixel analysis cannot?
2. Which components of difference can coarsening affect?
3. Is it necessary for the fine-resolution extent to be square in order to implement the coarsening algorithm?

References

Kuzera, K., & Pontius Jr, R. G. (2008). Importance of matrix construction for multiple-resolution categorical map comparison. *GIScience & Remote Sensing, 45*, 249–274. https://doi.org/10.2747/1548-1603.45.3.249.

Pontius Jr, R. G., & Cheuk, M. L. (2006). A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. *International Journal of Geographical Information Science, 20*, 1–30. https://doi.org/10.1080/13658810500391024.

Pontius Jr, R. G., & Connors, J. (2009). Range of categorical associations for comparison of maps with mixed pixels. *Photogrammetric Engineering and Remote Sensing, 75*, 963–969.

Pontius Jr, R. G., & Santacruz, A. (2015). *diffeR: Metrics of difference for comparing pairs of maps*. https://cran.r-project.org/web/packages/diffeR.

Pontius Jr, R. G., Walker, R., Yao-Kumah, R., Arima, E., Aldrich, S., Caldas, M., & Vergara, D. (2007). Accuracy assessment for a simulation model of Amazonian deforestation. *Annals of the Association of American Geographers, 97*, 677–695. https://doi.org/10.1111/j.1467-8306.2007.00577.x.

Pontius Jr, R. G., Boersma, W., Castella, J.-C., Clarke, K., de Nijs, T., Dietzel, C., Duan, Z., Fotsing, E., Goldstein, N., Kok, K., Koomen, E., Lippitt, C. D., McConnell, W., Mohd Sood, A., Pijanowski, B., Pithadia, S., Sweeney, S., Trung, T. N., Veldkamp, A. T., & Verburg, P. H. (2008). Comparing the input, output, and validation maps for several models of land change. *The Annals of Regional Science, 42*, 11–37. https://doi.org/10.1007/s00168-007-0138-2.

Chapter 7
Application to Categorical Temporal Change

Abstract This chapter applies the concepts of Chap. 4 to characterize land change during two time intervals 1971–1985 and 1985–1999 in a region of Massachusetts, USA. The case study has four categories: Built, Barren, Forest, and Water. A contingency table for each time interval are the inputs to the analysis. PontiusMatrix42.xlsx produces results and graphics similar to Intensity Analysis, which is a popular approach to analyze a contingency table. Multiple resolution analysis shows how Allocation differences vary across space.

Keywords Contingency table · Category · Land change · PontiusMatrix

7.1 Text

Figure 7.1 show maps consisting of pixels that are each 30 m by 30 m. Each pixel is one of four categories: Built, Barren, Forest, and Water. There are three time points: 1971, 1985, and 1999. An overlay of pairs of maps shows the changes during two time intervals: 1971–1985 and 1985–1999. Each patch of change consists of the loss of one category and the gain of a different category. It is helpful to see maps of losses and gains distinguished from persistence, especially when change constitutes a minority of the extent. Figure 7.1 shows also a contingency table that summarizes the number of pixels that derive from the maps. The table's rows are the categories at the start of each time interval while the table's columns are the categories at the end of each time interval. Numbers before the colon report sizes during 1971–1985 while numbers after the colon report sizes during 1985–1999. For example, the size of the transition from Barren to Built is 168 during 1971–1985 and is 138 during 1985–1999. Barren transitions to Built during the first time interval, then Barren transitions to both Built and Forest during the second time interval. Forest transitions to Built and Barren during both time intervals. The Sum column at the right shows that the categories at 1971 are in order of size: Forest, Barren, Built, and Water. The categories at 1985 are in order of size: Forest, Built, Barren, and Water.

The bottom of Fig. 7.1 shows components of overall difference at multiple resolutions with 1971–1985 on the left and 1985–1999 on the right. The coarsening of resolution cannot influence the Quantity components. The coarsening can shrink the

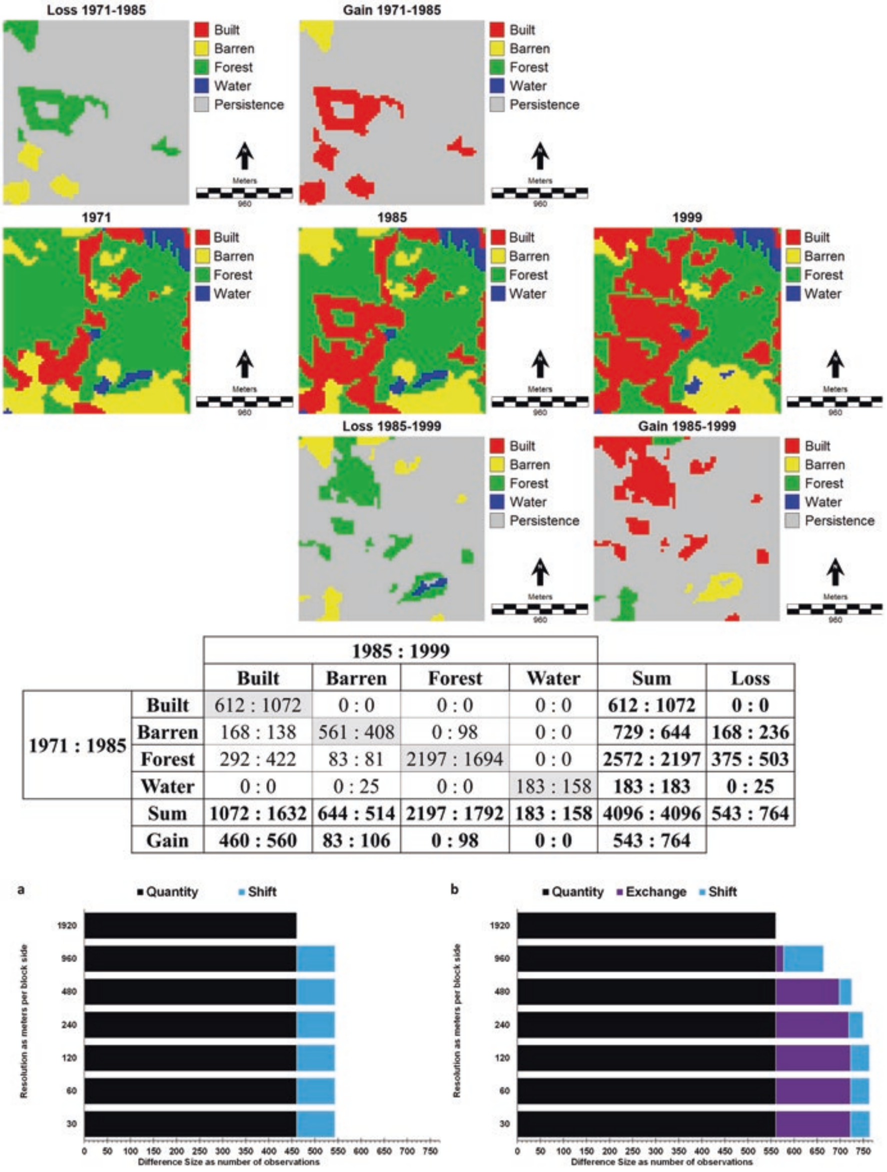
R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1_7

sum of Exchange and Shift, depending on whether a category at the fine resolution experiences simultaneous loss and gain in a coarser resolution block. The 960-m resolution has four blocks, which stratify the extent into four quadrants. The maps at the top of Fig. 7.1 show that Barren is the source of the Shift during 1971–1985, as Barren transitions to Built in the southwest quadrant and Forest transitions to Barren the northeast quadrant. Barren does not experience simultaneous loss and gain in any quadrant, thus the Barren's Shift at the 30-m resolution is identical to its Shift at the 960-m resolution. At the 1920-m resolution, Exchange and Shift are zero because Allocation difference does not exist in a single block. During 1985–1999, Barren and Forest experience both loss and gain, thus generate Allocation difference, which is the sum of Exchange and Shift. Allocation difference is 204 observations at the raw data's 30-m resolution. The coarsening to 960 m causes Allocation difference to shrink by half. This means that half of the Allocation difference occurs within quadrants while the other half occurs across quadrants. Figure 7.1b illustrates how coarser resolutions cannot cause the sum of Exchange and Shift to grow but can cause either Exchange or Shift to grow, as the Shift is larger at the 960-m resolution than at the 480-m resolution.

We use the methods of Chap. 4 to analyze the contingency table for each time interval. Figure 7.2 shows results for 1971–1985 while Fig. 7.3 shows results for 1985–1999. The results and graphics derive from PontiusMatrix42.xlsx (Pontius Jr 2020).

Parts a–c on the left side of Figs. 7.2 and 7.3 show sizes in the same units that appear in the tables. Part a of each figure shows the sizes of the entries in the contingency table. The label Per is an abbreviation for Persistence, which appears in the table's diagonal entries. The sum of gains must equal the sum of losses, as the bottom of part a shows. Part b of each figure shows the Venn diagrams that compare each category's start time with its end time. The intersection of the two times is the persistence in the middle of each Venn diagram. The Quantity component of change resides at one end of each segmented bar. Built experiences Quantity gain during both time intervals. Barren and Forest experience Quantity loss during both time intervals. Other segments of a category's Venn diagram show the category's Exchange and Shift. Barren experiences Shift during the first time interval, as Barren transitions to Built while Forest transitions to Barren. During the second time interval, Barren experiences also Exchange, as Barren transitions to Forest while Forest transitions to Barren. Part c of Figs. 7.2 and 7.3 show the sizes of the changes. The extent bar shows the size of change over all categories. The other bars show size of change by category. Each category's Quantity component has a label of Loss or Gain. A label of Loss indicates the category lost more than it gained. A label of Gain indicates the category gained more than it lost.

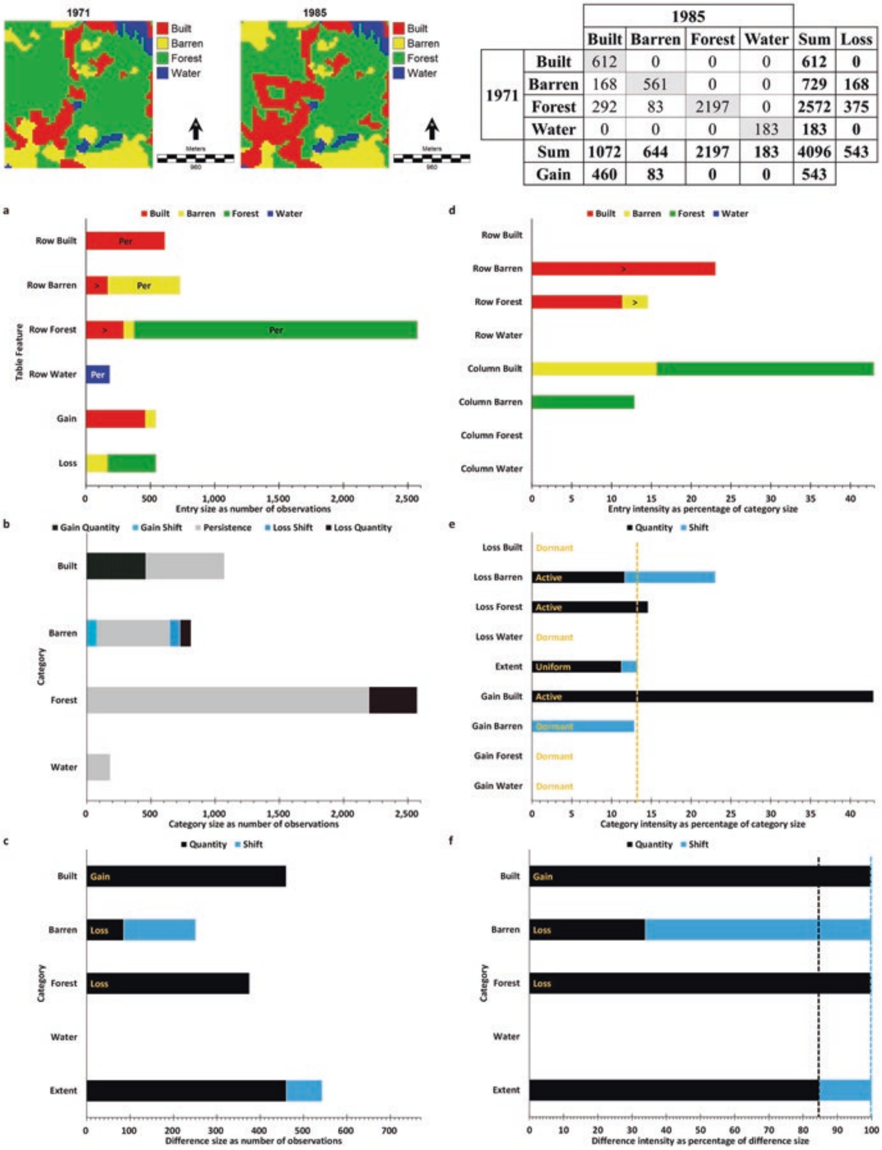
Parts d–e on the right side of Figs. 7.2 and 7.3 show intensities, which are ratios expressed as percentages. Part d of Figs. 7.2 and 7.3 shows the transition intensities, where a transition is an off-diagonal entry in the contingency table. Equation 4.17 computes the transition intensity as the size of the transition divided by the start size of the losing category. The resulting ratio is known also as Markov proportion, which is a popular way to express a transition (Varga et al. 2019). When a particular



		1985 : 1999					
		Built	**Barren**	**Forest**	**Water**	**Sum**	**Loss**
1971 : 1985	**Built**	612 : 1072	0 : 0	0 : 0	0 : 0	**612 : 1072**	**0 : 0**
	Barren	168 : 138	561 : 408	0 : 98	0 : 0	**729 : 644**	**168 : 236**
	Forest	292 : 422	83 : 81	2197 : 1694	0 : 0	**2572 : 2197**	**375 : 503**
	Water	0 : 0	0 : 25	0 : 0	183 : 158	**183 : 183**	**0 : 25**
	Sum	**1072 : 1632**	**644 : 514**	**2197 : 1792**	**183 : 158**	**4096 : 4096**	**543 : 764**
	Gain	**460 : 560**	**83 : 106**	**0 : 98**	**0 : 0**	**543 : 764**	

Fig. 7.1 Land change during two time intervals: (**a**) 1971–1985 and (**b**) 1985–1999

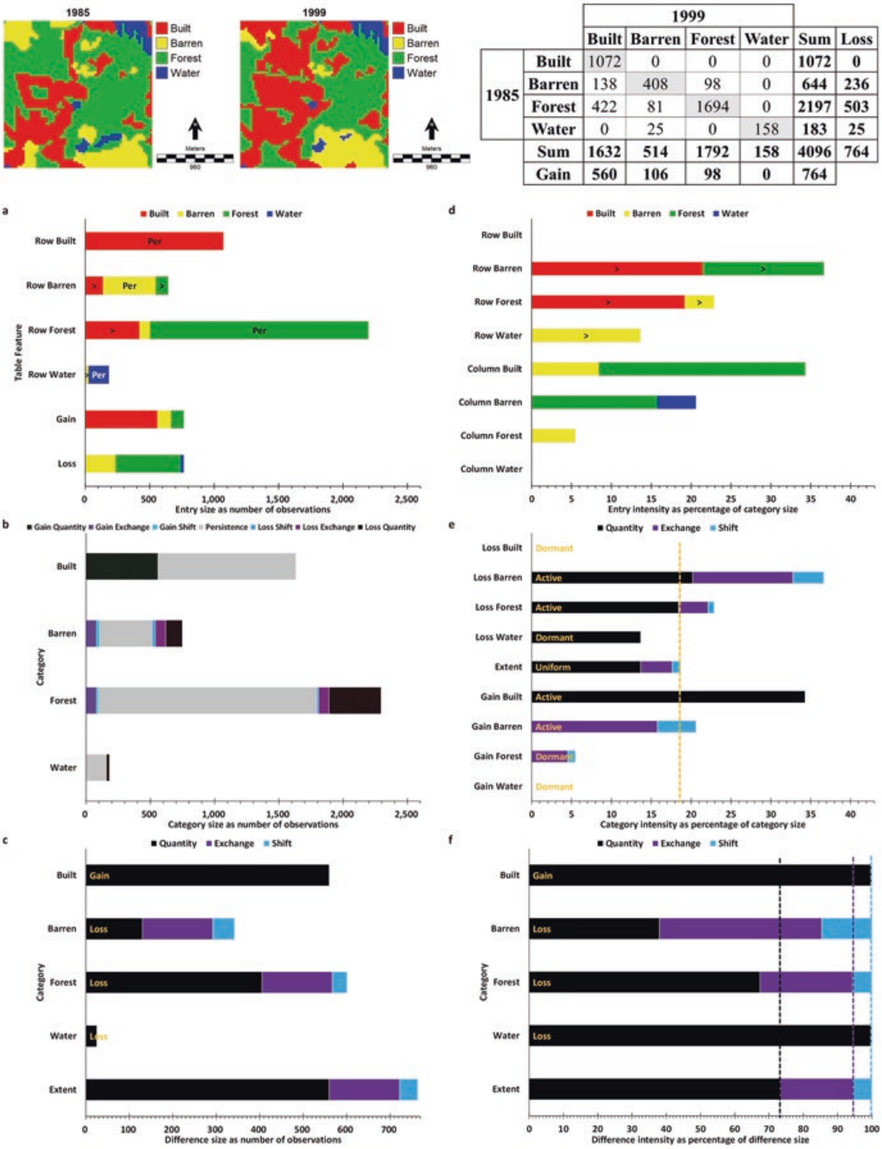
category gains, if the gain's transition intensity from all the losing categories were equal, then they would equal the result from Eq. 4.41, which is known as the uniform transition intensity. The results from Eq. 4.17 relative to the result from Eq. 4.41 determines the labels on the bars in part d of Figs. 7.2 and 7.3. A label of > indicates that a particular transition intensity is greater than the gaining category's



		1985					
		Built	Barren	Forest	Water	Sum	Loss
1971	Built	612	0	0	0	612	0
	Barren	168	561	0	0	729	168
	Forest	292	83	2197	0	2572	375
	Water	0	0	0	183	183	0
	Sum	1072	644	2197	183	4096	543
	Gain	460	83	0	0	543	

Fig. 7.2 Results for change during 1971–1985

uniform transition intensity, in which case we say the gaining category targets the losing category. A label of = indicates that a particular transition intensity equals the gaining category's uniform transition intensity. If a particular transition intensity is less than gaining category's uniform transition intensity, then the transition receives no label and we say the gaining category avoids the losing category. We must



		1999					
		Built	Barren	Forest	Water	Sum	Loss
1985	Built	1072	0	0	0	1072	0
	Barren	138	408	98	0	644	236
	Forest	422	81	1694	0	2197	503
	Water	0	25	0	158	183	25
	Sum	1632	514	1792	158	4096	764
	Gain	560	106	98	0	764	

Fig. 7.3 Results for change during 1985–1999

compare the transition intensities for each gaining category in part d of the figures. The red segments in Fig. 7.2d indicate that Built's gain targets Barren while avoids both Forest and Water during the first time interval. The red segments in Fig. 7.2a, d highlight patterns that allow us to test evidence for hypothesized processes. Figures 7.2a shows that Built gains more from Forest than from Barren in terms of size, but that does not imply that builders demonstrate a preference to build on Forest. The size of Forest is larger than the size of Barren at 1971, thus more Forest than Barren is available to builders at 1971. The intensities in Fig. 7.2d show that Built gains more intensively from Barren than from Forest. In fact, evidence supports a hypothesis that builders target Barren and avoid Forest. The red segments in Fig. 7.3d shows that Built's gain targets both Barren and Forest while avoids Water during the second time interval. The yellow segments in Fig. 7.3d shows that Barren's gain targets both Forest and Water during the second time interval. Part e of Figs. 7.2 and 7.3 shows each category's loss intensity and gain intensity. A category's loss intensity is its loss expressed as a percent of its start size. A category's gain intensity is its gain as a percent of its end size. The extent bar indicates the overall change as a percentage of the extent, which accelerates from 13% during the first time interval to 18% during the second time interval. If a category's loss intensity or gain intensity is greater than the extent's intensity, then the category's loss or gain is active during the time interval. During the first time interval, active changes are Barren's loss, Forest's loss, and Built's gain. During the second time interval, active changes are Barren's loss, Forest's loss, Built's gain, and Barren's gain. Part f of Figs. 7.2 and 7.3 show the components' intensities. The dashed lines show how the components contribute to the overall difference in the extent. The Quantity component accounts for most of the extent's difference during both time intervals. In contrast, the Quantity component accounts less than 40% of the Barren's difference during both time intervals.

The methods that produced parts d and e of Figs. 7.2 and 7.3 follow the same logic of Intensity Analysis. Intensity Analysis is a framework to analyze a series of contingency tables (Aldwaik and Pontius Jr 2012, 2013; Pontius Jr et al. 2013). Many applications of Intensity Analysis have examined land change across various time intervals (Huang et al. 2018; Quan et al. 2020). Intensity Analysis has three levels: interval, category, and transition. The interval level compares the speed of change among time intervals. The durations of both time intervals in this chapter's example are 14 years, thus it makes sense to compare Fig. 7.2 with Fig. 7.3. If the durations of the time intervals were not identical, then it would be necessary to divide the change during each time interval by the duration of the time interval in order to compare across time intervals. The intensity.analysis package in R performs the calculations necessary to compare several time intervals that have various durations (Pontius Jr and Khallaghi 2019). Part e of Figs. 7.2 and 7.3 show results for Intensity Analysis' category level. Part d of Figs. 7.2 and 7.3 show results for Intensity Analysis' transition level. Part d shows the transition intensities for every gaining category, which is a more efficient graphical display than in the literature that has applied Intensity Analysis. The Intensity Analysis literature so far has usually presented a graphic for each gaining category at the transition level, which causes many graphics that can be overwhelming for readers.

7.2 Discussion Questions

1. What parts of this chapter's figures show the overall change in the extent?
2. What are the meanings of dormant and active?
3. What are the meanings of avoid and target?
4. What is the evidence concerning whether builders prefer to gain from the existing Barren more than from the existing Forest during each of the time intervals that this chapter analyzes?
5. Where are Markov proportions in this chapter's figures?
6. What insights does multiple resolution analysis give to this chapter's analysis of temporal change?

References

Aldwaik, S. Z., & Pontius Jr, R. G. (2012). Intensity analysis to unify measurements of size and stationarity of land changes by interval, category, and transition. *Landscape and Urban Planning, 106*, 103–114. https://doi.org/10.1016/j.landurbplan.2012.02.010.

Aldwaik, S. Z., & Pontius Jr, R. G. (2013). Map errors that could account for deviations from a uniform intensity of land change. *International Journal of Geographical Information Science, 27*, 1717–1739. https://doi.org/10.1080/13658816.2013.787618.

Huang, B., Huang, J., Pontius Jr, R. G., & Tu, Z. (2018). Comparison of Intensity Analysis and the land use dynamic degrees to measure land changes outside versus inside the coastal zone of Longhai, China. *Ecological Indicators, 89*, 336–347. https://doi.org/10.1016/j.ecolind.2017.12.057.

Pontius Jr, R. G. (2020). *PontiusMatrix42.xlsx*. http://www.clarku.edu/~rpontius

Pontius Jr, R. G., & Khallaghi, S. (2019). *Intensity.analysis*. https://cran.r-project.org/web/packages/intensity.analysis

Pontius Jr, R. G., Gao, Y., Giner, N., Kohyama, T., Osaki, M., & Hirose, K. (2013). Design and interpretation of intensity analysis illustrated by land change in Central Kalimantan, Indonesia. *Land, 2*, 351–369. https://doi.org/10.3390/land2030351.

Quan, B., Pontius Jr, R. G., & Song, H. (2020). Intensity Analysis to communicate land change during three time intervals in two regions of Quanzhou City, China. *GIScience & Remote Sensing, 57*, 21–36. https://doi.org/10.1080/15481603.2019.1658420.

Varga, O. G., Pontius Jr, R. G., Singh, S. K., & Szabó, S. (2019). Intensity Analysis and the figure of Merit's components for assessment of a Cellular Automata – Markov simulation model. *Ecological Indicators, 101*, 933–942. https://doi.org/10.1016/j.ecolind.2019.01.057.

Chapter 8
Interval Variable Versus Interval Variable

Abstract This chapter gives methods to compare **Y** versus **X** when both variables indicate a phenomenon with the same units on an interval scale. The analysis' begins with a visual examination of a square scatter plot of where each observation is a point (*X*,*Y*) plotted relative to the line $Y = X$. Mean Deviation is the average **Y** minus the average **X**. Mean Absolute Deviation is the average vertical distance between the $Y = X$ line and the points in the scatter plot. Mean Absolute Deviation is the sum of two components: Quantity and Allocation. Correlation is an index on the continuous interval [−1,1] concerning the strength and sign of a linear relationship between **Y** and **X**. The slope of the least squares line indicates the change in **Y** for each increment increase in **X**. Each of those four metrics measures a distinct concept. Relevant software includes the diffeR package available at https://cran.r-project.org/web/packages/diffeR/index.html (Pontius Jr and Santacruz 2015).

Keywords Allocation · Correlation · Interval · Mean Deviation · Mean Absolute Deviation · Slope

8.1 Text

This chapter describes how to compare **X** with **Y** when both variables show the same units of a phenomenon on an interval scale (Pontius Jr et al. 2008). Interval scales can measure continuous phenomena, such as temperature, or countable phenomena, such as number of people. If **X** and **Y** are on the same interval scale, then each observation's deviation has a clear interpretation. For example, if *X* is 3 and *Y* is 1 for an observation on an interval scale, then the deviation of *Y* minus *X* is −2. If a variable is a numerical code, then the variable is not necessarily an interval scale. For example, if 1 indicates Low, 2 indicates Medium, and 3 indicates High, then the variable is not interval because the difference between 2 and 1 does not necessarily have the same interpretation as twice the difference between 3 and 1. An intelligent first step to compare two variables that are on the same interval scale is to examine a square scatter plot of **Y** versus **X** where each observation is a point (*X*,*Y*) and the plot includes the $Y = X$ line as a reference. Visual examination relative to the $Y = X$ line can reveal patterns that summary metrics fail to indicate. Visual examination

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1_8

can also miss some important relationships between **X** and **Y**, especially when many observation points in the plot are piled on top of each other. This chapter gives a collection of helpful metrics to quantify the relationship between **X** and **Y** in ways that complement a visual assessment.

This chapter illustrates its concepts by comparing **X** to nine series for **Y**, where the series have the names A, B, C, …, I. Figure 8.1 shows scatter plots for the nine series, where each series has four observations. Figure 8.1a–c shows two series per plot, where each series name appears as a letter in each of the four observations. Figure 8.1d shows series G, H, and I. All plots have the same **X** values, which are 8, 9, 11, and 12. Table 8.1 gives the deviation from the $Y = X$ line for each observation, so readers can see the organization of the data and work through the calculations. Series A and B are identical concerning their deviations, meaning two are −4 and two are 4; however, the pairing of the four deviations with the four **X** values varies between A and B. Positive deviations are paired with larger **X** values in series A. Negative deviations are paired with larger **X** values in series B. Similarly, series C and D have the same four deviations but C and D differ in how the four deviations are paired with the **X** values. The same relationship concerning the pairing exists for E and F, and also for G and H. All the deviations are negative 4 for series I.

Table 8.2 gives the mathematical notation that this chapter uses to compare two variables that are on the same interval scale. Equation 8.1 computes the average **X** while Eq. 8.2 computes the average **Y**. Equation 8.3 gives the deviation for each observation. Equation 8.4 defines the Mean Deviation (MD), which is the average **Y** minus the average **X**. Some authors use the word "bias" to refer to Mean Deviation, which can be negative, zero, or positive (Willmott and Matsuura 2005, 2006). Equation 8.4 shows that MD ignores how each **X** observation is paired with each **Y** observation, as MD does not require knowledge of each D_i. Equation 8.5 gives the Mean Absolute Deviation (MAD), which is the average vertical distance between the $Y = X$ line and each point in the scatter plot. MAD equals also the average horizontal distance between the $Y = X$ line and each point in the scatter plot. Equation 8.6 shows that MAD considers how each **X** observation is paired with each **Y** observation, as MAD requires knowledge of each D_i. MAD is the sum of two components: Quantity and Allocation. Equation 8.6 gives the Quantity component, which

Table 8.1 Deviations of Y minus X for each of four observations for each of nine series for **Y**

Y series	$X = 8$	$X = 9$	$X = 11$	$X = 12$
A	−4	−4	4	4
B	4	4	−4	−4
C	−7	−1	1	7
D	7	1	−1	−7
E	−7	−7	1	1
F	1	1	−7	−7
G	−7	−7	−1	−1
H	−1	−1	−7	−7
I	−4	−4	−4	−4

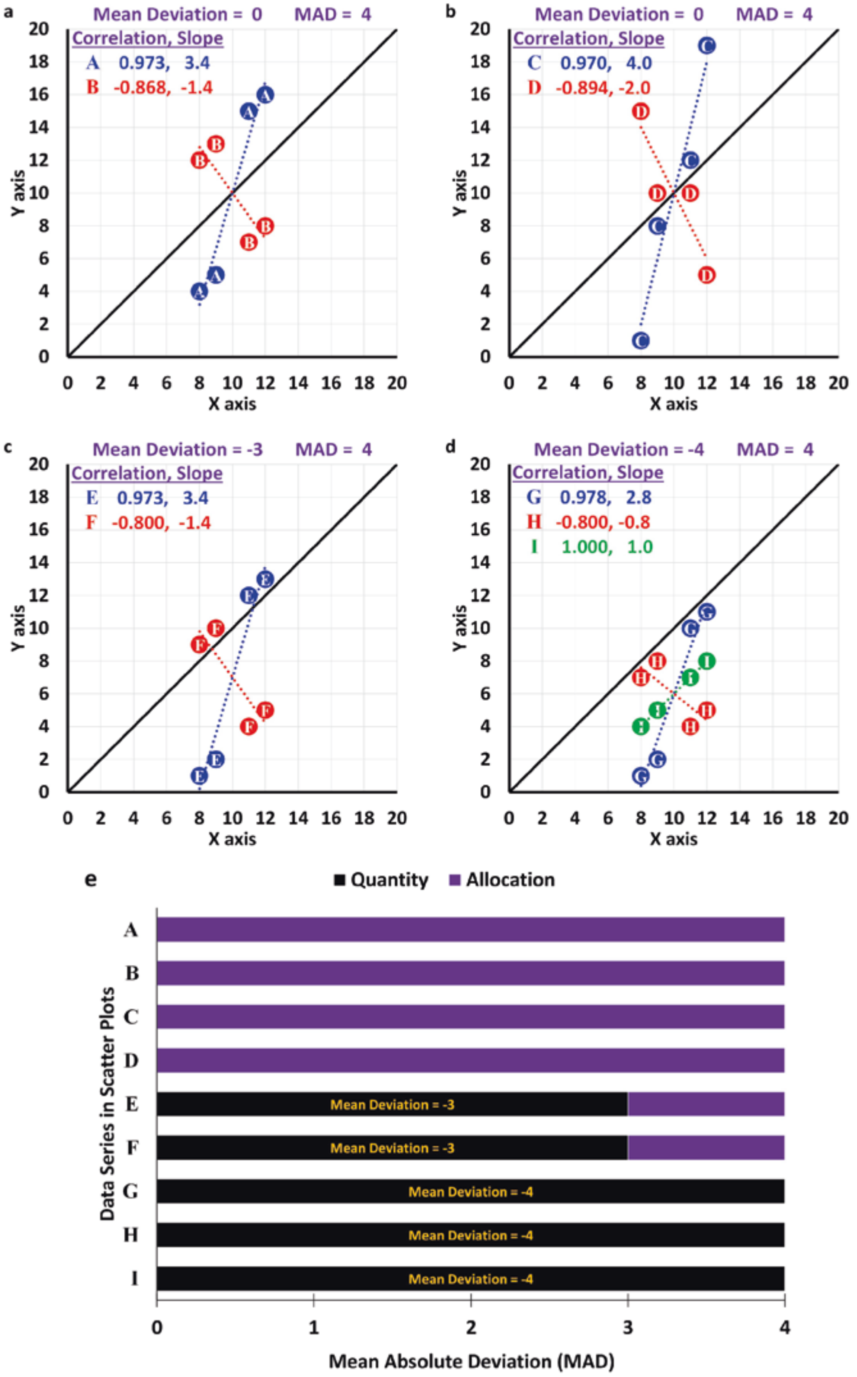


Fig. 8.1 Example data to compare variables that show the same phenomenon on an interval scale

Table 8.2 Notation to compare two variables that show the same interval phenomenon

Notation	Meaning
β	Slope of the least squares line
D_i	Deviation for observation i
i	Index for observation where $i = 1, 2, \ldots N$
N	Number of observations
r	Pearson's correlation coefficient
X	Arbitrary value for the independent variable
X_i	Value of **X** for observation i
$\overline{X}$	Mean of **X**
Y	Arbitrary value for the dependent variable
Y_i	Value of **Y** for observation i
$\overline{Y}$	Mean of **Y**

is the absolute value of MD. Equation 8.7 gives the Allocation Deviation, which is MAD minus the Quantity Deviation.

$$\overline{X} = \sum_{i=1}^{N} X_i / N \tag{8.1}$$

$$\overline{Y} = \sum_{i=1}^{N} Y_i / N \tag{8.2}$$

$$D_i = Y_i - X_i \tag{8.3}$$

$$\text{Mean Deviation} = \overline{Y} - \overline{X} \tag{8.4}$$

$$\text{Mean Absolute Deviation} = \sum_{i=1}^{N} \left| D_i \right| / N \tag{8.5}$$

$$\text{Quantity Deviation} = \left| \overline{Y} - \overline{X} \right| = \left| \sum_{i=1}^{N} D_i \right| / N \tag{8.6}$$

$$\text{Allocation Deviation} = \left(\sum_{i=1}^{N} \left| D_i \right| - \left| \sum_{i=1}^{N} D_i \right| \right) / N \tag{8.7}$$

Figure 8.1 shows that Mean Deviation is zero for series A, B, C, and D. Mean Deviation is negative 3 for series E and F. Mean Deviation is negative 4 for series G, H and I. Mean Absolute Deviation is 4 for all series A-I. Mean Deviation (MD) is frequently the first helpful metric to understand when analyzing a scatter plot. If **X** is the truth and **Y** is a diagnosis, then MD reveals the size of the average error, including whether the average diagnosis is less than, equal to, or greater than the average truth. If **X** is the start time and **Y** is the end time of a phenomenon, then the MD reveals the size of the phenomenon's average net change, which can be negative, zero, or positive. MD ignores the distances between the $Y = X$ line and the points in the scatter plot. Mean Absolute Deviation (MAD) is the average distance between the $Y = X$ line and the points. If **X** is the truth and **Y** is a diagnosis, then MAD is the average absolute error. If **X** is the start time and **Y** is the end time, then MAD is the size of the average absolute change. The units of both MD and MAD are identical to the units of **X** and **Y**, which makes interpretation straightforward because we can interpret MD and MAD in terms of the phenomenon that **X** and **Y** describe.

Equations 8.4, 8.5, 8.6 and 8.7 imply that MD $\leq$ |MD| = Quantity Deviation $\leq$ (Quantity Deviation + Allocation Deviation) = MAD. Figure 8.1e shows for each series how MAD is the sum of its two components: Quantity and Allocation. The Allocation Deviation is positive if and only if the scatter plot has at least one point above the $Y = X$ line and at least one point below the $Y = X$ line. In other words, the Allocation Deviation is positive if and only if at least one D_i is positive and at least one D_i is negative. If that is not the case, then the non-zero deviations contribute exclusively to the Quantity Deviation. The concepts of the components for the interval variable are analogous to the concepts for a binary variable, for which Exchange is positive if and only if a case has both False Alarms and Misses. If False Alarms are zero or Misses are zero, then all the differences contribute exclusively to the Quantity component for a binary variable.

MAD reveals information concerning how **X** relates to **Y** by summarizing the values for D_i. But MAD does not consider the pairings between each D_i and each X_i. Scatter plots show how each D_i pairs with each X_i. The pairings reveal the association between **X** and **Y**. If the scatter plot is square and includes the $Y = X$ line, then the plot usually reveals the association clearly. Several metrics exist to measure various aspects of the association. The remainder of this chapter examines popular metrics and their interpretations.

Equations 8.8 and 8.9 give definitions of variance for **X** and **Y**, which are necessary to interpret some measures of association. Equation 8.10 gives Pearson's correlation coefficient, which describes the strength and sign of a linear relationship between **X** and **Y**. Other measures of correlation exist, such as Spearman's rank correlation. If a piece of literature does not specify the type of correlation, then it is likely to be Pearson's correlation. Pearson's correlation ranges from −1 to 1. Pearson's correlation is −1 if and only if a perfectly linear negatively sloped relationship exists between **X** and **Y**. Pearson's correlation is 1 if and only if a perfectly linear positively sloped relationship exists between **X** and **Y**. If Pearson's correlation is zero, then neither an increasing nor decreasing linear relationship exists between **X** and **Y**. If all the **X** values are identical or all the **Y** values are identical, then the Pearson's correlation coefficient is 0/0, which is undefined. Equation 8.11 computes the square of Pearson's correlation, which measures the strength of a linear relationship on a scale from 0 to 1, where 0 means no linear relationship and 1 means a perfect linear relationship. Equation 8.11 is known as R-squared and also as the Coefficient of Determination for this examination of a linear relationship. The interpretation of R-squared is the proportion of the variance in **Y** that a relationship with **X** explains. R-squared can be confusing in some literature because R-squared is a general term that might indicate the strength of a non-linear relationship between **X** and **Y**. Authors must clarify the mathematical form of the relationship between **X** and **Y** when reporting R-squared. One reason why I recommend that authors report Correlation rather than R-squared is to avoid possible confusion concerning the mathematical relationship that R-squared measures. Pearson's correlation refers specifically to a linear association. If readers know Correlation then they can compute R-squared but not vice versa, which is another reason why I recommend that authors report Correlation rather than R-squared. R-squared fails to indicate whether a relationship is increasing or decreasing, which is a third reason why I recommend Correlation rather than R-squared. Neither R-squared nor Correlation indicate the steepness of a linear relationship between **X** and **Y**. Equation 8.12 gives the slope of the fitted least squares line, which is a popular type of trend line. The least squares line derives from ordinary least squares linear regression. The slope of the least squares line describes the deviation in **Y** for each increment increase in **X**. Slope shows whether a linear relationship is negative, zero, or positive, just as Correlation shows. Slope shows also the steepness of the relationship, which Correlation does not indicate. It is helpful to know whether Slope is less than one or greater than one because the slope of the $Y = X$ line is one. Slope greater than one indicates that larger D_i values tend to be associated with larger X_i values. Slope less than one indicates that smaller D_i values tend to be associated with larger X_i values. Slope equal to one indicates D_i values are not associated linearly with X_i values. Correlation indicates the strength and sign of a linear relationship but Correlation fails to indicate the steepness of the Slope of the linear relationship. Slope expresses how **Y** changes with each increment increase in **X**, but Slope does not measure the strength of the linear relationship. Equation 8.13 gives the Intercept of the least squares line, which is necessary to report the equation for the least squared line. Equation 8.14 gives the equation of the least squares line, which can be helpful to make predictions for new observations of the independent variable X. If the equation of the line is not essential for your particular application, then I recommend you refrain from reporting Intercept because I have seen people misinterpret Intercept as the Mean Deviation, which Intercept is not. For example, series F and H illustrate cases where Mean Deviation is negative while Intercept is positive.

$$\text{Variance in } \mathbf{X} = \sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2 / N \tag{8.8}$$

$$\text{Variance in } \mathbf{Y} = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2 / N \tag{8.9}$$

$$\text{Correlation} = r = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\left[\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2\right]\left[\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2\right]}} \tag{8.10}$$

$$r^2 = \frac{\left[\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)\right]^2}{\left[\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2\right]\left[\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2\right]} \tag{8.11}$$

$$\text{Slope} = \beta = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2} \tag{8.12}$$

$$\text{Intercept} = \overline{Y} - \beta\overline{X} \tag{8.13}$$

$$Y = \beta X + \text{Intercept} \tag{8.14}$$

Figure 8.1 demonstrates how the pairings between D_i and X_i influence both Correlation and Slope. Larger D_i are associated with larger X_i in series A, C, E, and G; thus, Slope is greater than one. Smaller D_i are associated with larger X_i in series B, D, F, and H; thus, Slope is less than one. The four D_i are identical in series I; thus, the Slope is one. Series I demonstrates also that Correlation is one when the points form a straight line.

The collection of MD, MAD, Correlation, and Slope indicate various aspects concerning how the points in the scatter plot are arranged relative to the $Y = X$ line, which is why the collection of those four metrics facilitates interpretation. MD = 0 if and only if the average **Y** equals the average **X**. MAD = 0 if and only if all the points in the scatter plot are on the $Y = X$ line. If all the points are on the $Y = X$ line and Variance in **X** is positive, then MD = 0. Correlation = 1 if and only if there is a perfectly positive linear relationship between **X** and **Y**. Slope = 1 if and only if the least squares line is parallel to the $Y = X$ line.

A metric is symmetric when the reversal of **X** and **Y** does not influence the metric. In other words, a metric is symmetric when the analysis of **Y** versus **X** produces the same metric as the analysis of **X** versus **Y**. Symmetric metrics include MAD, the Quantity Deviation, the Allocation Deviation, and Correlation. MD is not symmetric, which allows MD to reveal whether average **Y** or average **X** is larger. Slope is not symmetric, meaning the fitted linear relationship between **X** and **Y** depends on how we assign the two variables to **X** and **Y**. If β is the Slope for one possible assignment, then the Slope for the other possible assignment is not necessarily β or $1/\beta$. Thus, we must think carefully concerning which variable to assign as **X** and which to assign as **Y** when computing Slope. The convention is for **X** to be the independent variable upon which **Y** might depend, thus **Y** is the dependent variable. For example, when comparing the truth to a diagnosis, **X** would be truth and **Y** would be the diagnosis because the truth exists independently from the diagnosis, while the diagnosis might depend on the truth. When comparing a start time to an end time, **X** would be start time and **Y** would be the end time because the start time exists independently from the end time, while the end time might depend on the start time.

Several of this chapter's metrics are so common that many software packages compute them. However, Quantity Deviation and Allocation Deviation are less common. The diffeR package in R reads raster maps to compute Quantity and Allocation deviations (Pontius Jr and Santacruz, 2015). The package can compute the metrics at multiple spatial resolutions, which shows how Allocation Deviation shrinks when coarser resolutions create larger blocks that contain both positive and negative deviations.

8.2 Discussion Questions

1. Which equation(s) in this chapter measure mean distance between the $Y = X$ line and the (X,Y) points?
2. If Correlation equals one, then is it possible for Slope to be zero?
3. If Correlation equals one, then how close to zero could Slope be?
4. What are three advantages of reporting Correlation rather than R-squared?
5. If the slope of the least squares line equals one, then what is the relationship between Mean Deviation and the line's Intercept?
6. If Mean Deviation equals the Intercept of the least squares line, then the line's slope equals what?
7. What influential decision must you make before you compute a metric that is not symmetric?
8. What advice would you give to a scientist who reports only R-squared to describe the association between **X** and **Y** when both variables describe the same phenomenon?

References

Pontius Jr, R. G., & Santacruz, A. (2015). *diffeR: Metrics of difference for comparing pairs of maps*. https://cran.r-project.org/web/packages/diffeR

Pontius Jr, R. G., Thontteh, O., & Chen, H. (2008). Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics, 15*, 111–142. https://doi.org/10.1007/s10651-007-0043-y.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*, 79–82. https://doi.org/10.3354/cr030079.

Willmott, C. J., & Matsuura, K. (2006). On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International Journal of Geographical Information Science, 20*, 89–102. https://doi.org/10.1080/13658810500286976.

Chapter 9
Application to Interval Temporal Change

Abstract This chapter applies the concepts of the previous chapter to compare maps concerning a continuous phenomenon, specifically sea surface temperature. The **X** variable derives from February 1982. We consider two **Y** variables at August 1982 and February 2010. This chapter gives equations to stratify the analysis. The global data have two strata: the northern hemisphere and the southern hemisphere. The metrics, scatter plots, and maps each reveal information that the other two forms of presentation do not reveal. Mean Deviation reveals global cooling during the half-year interval and global warming during the 28-year interval. The components of Mean Absolute Deviation quantify how most of the change between seasons is Allocation deviation across hemispheres, while most of the change between years is Quantity deviation. Correlation and the scatter plots show a stronger linear association between times during the 28-year interval than during the half-year interval. The maps show the spatial distribution of the temperature deviations, which the scatter plots and most of the metrics fail to indicate.

Keywords Allocation · Correlation · Change · Interval · Mean Absolute Deviation · Slope

9.1 Text

Some places of the globe routinely experience temperature changes by dozens of degrees Celsius across seasons, so a couple of degrees global warming across years might seem unthreatening to some people. Science needs methods to communicate how the change across years differs conceptually from the change across seasons. This chapter illustrates helpful ways to communicate temporal change for an interval phenomenon, i.e. sea surface temperature.

Figure 9.1 shows the data and results for this case study. The maps at the top show sea surface temperature at three time points: February 1982, August 1982, and February 2010. Two strata differentiate the northern hemisphere from the southern hemisphere. The maps are in the Mollweide projection, which is an equal-area projection. Each pixel represents 100 square kilometers. The equator has the largest

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1_9

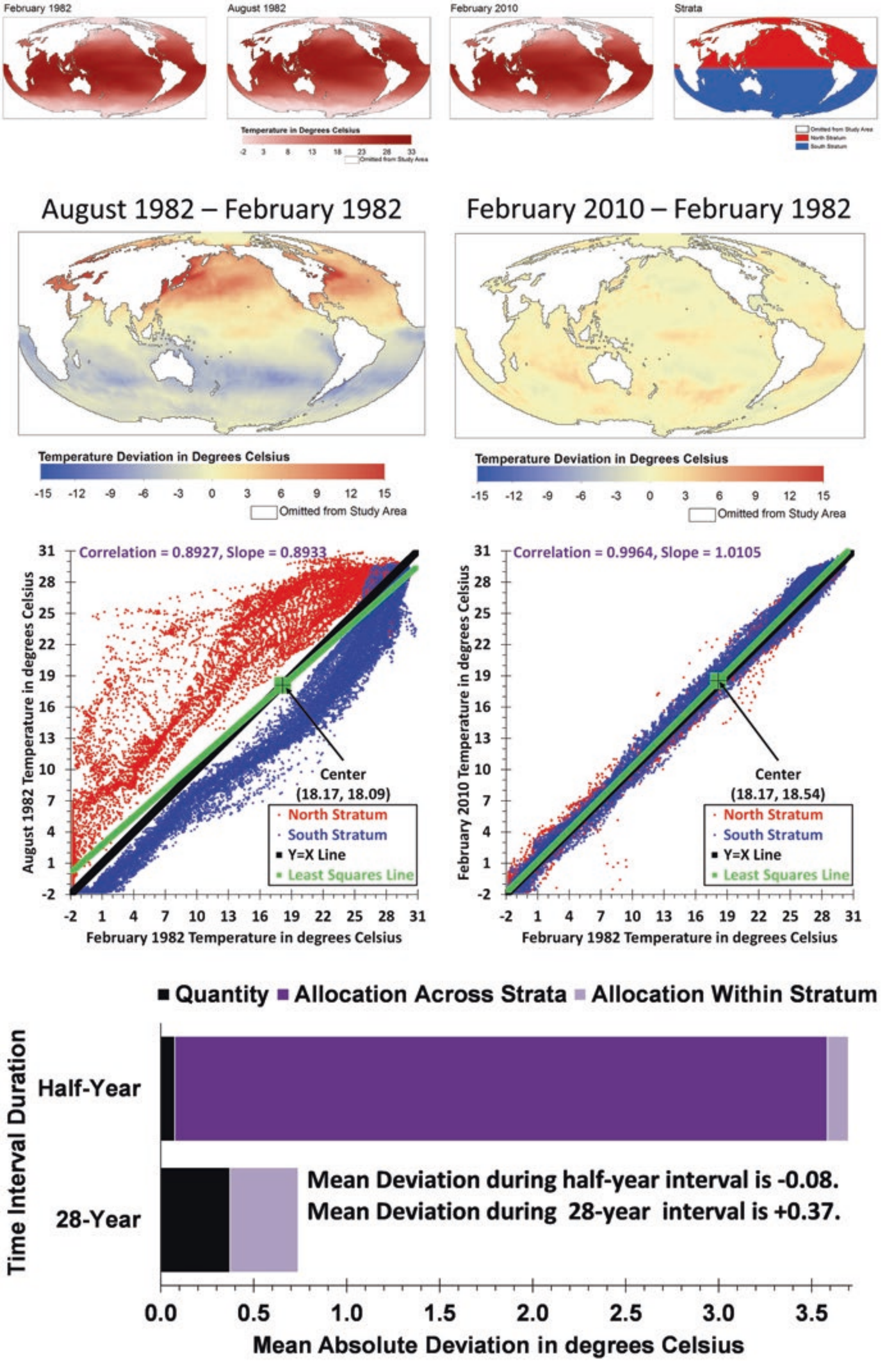


Fig. 9.1 Application of comparison of interval variables to temporal change

Table 9.1 Notation for interval variables when observations are grouped into strata

Notation	Meaning
β	Slope of least squares line
b	Index for a stratum
B	Number of strata
D_{b_i}	Deviation for observation *i* within stratum *b*
i	Index for observation within a stratum where $i = 1, 2, \ldots N_b$
N_b	Number of observations in stratum *b*
N	Number of observations
r	Pearson's correlation coefficient
X_{b_i}	Value of **X** for observation *i* within stratum *b*
Y_{b_i}	Value of **Y** for observation *i* within stratum *b*

number of pixels across any particular latitude. The number of pixels across a latitude shrinks as the latitude approaches each pole.

The maps at the three time points are difficult to differentiate visually, so Fig. 9.1 shows also maps of the temperature change from February 1982 to two time points: August 1982 and February 2010. The map of change during the half year in 1982 shows warming in the north and cooling in the south. The map of change during 28 years shows that more regions are warming than are cooling. The scatter plot below each map shows February 1982 on the horizontal *X* axis and the end time point on the vertical *Y* axis. Each pixel location in the maps generates a point in the scatter plot. Blue points derive from the southern hemisphere, while red points derive from the northern hemisphere. If a pixel's temperature does not change, then the pixel's point appears on the diagonal $Y = X$ line. If a pixel experiences warming, then its point appears above the diagonal line. If a pixel experiences cooling, then its point appears below the diagonal line. This chapter's equations measure the arrangement of the points relative to the diagonal line. Table 9.1 gives the mathematical notation. The equations express the same concepts as the previous chapter, with the added feature that this chapter considers how the observations are grouped into strata (Pontius Jr et al. 2008).

$$N = \sum_{b=1}^{B} N_b \tag{9.1}$$

$$D_{b_i} = Y_{b_i} - X_{b_i} \tag{9.2}$$

$$\text{Mean Deviation} = \sum_{b=1}^{B} \sum_{i=1}^{N_b} D_{b_i} \;/\; N \tag{9.3}$$

$$\text{Quantity Deviation} = \left| \sum_{b=1}^{B} \sum_{i=1}^{N_b} D_{b_i} \right| / \; N \tag{9.4}$$

$$\text{Allocation Across Strata Deviation} = \left(\sum_{b=1}^{B} \left| \sum_{i=1}^{N_b} D_{b_i} \right| - \left| \sum_{b=1}^{B} \sum_{i=1}^{N_b} D_{b_i} \right| \right) / N \tag{9.5}$$

$$\text{Allocation Within Stratum Deviation} = \left(\sum_{b=1}^{B} \sum_{i=1}^{N_b} \left| D_{b_i} \right| - \sum_{b=1}^{B} \left| \sum_{i=1}^{N_b} D_{b_i} \right| \right) / N \tag{9.6}$$

$$\text{Mean Absolute Deviation} = \sum_{b=1}^{B} \sum_{i=1}^{N_b} \left| D_{b_i} \right| / N \tag{9.7}$$

$$\bar{X} = \sum_{b=1}^{B} \sum_{i=1}^{N_b} X_{b_i} \; / \; N \tag{9.8}$$

$$\bar{Y} = \sum_{b=1}^{B} \sum_{i=1}^{N_b} Y_{b_i} \; / \; N \tag{9.9}$$

$$\text{Correlation} = r = \frac{\sum_{b=1}^{B} \sum_{i=1}^{N_b} \left(X_{b_i} - \bar{X} \right) \left(Y_{b_i} - \bar{Y} \right)}{\sqrt{\left[\sum_{b=1}^{B} \sum_{i=1}^{N_b} \left(X_{b_i} - \bar{X} \right)^2 \right] \left[\sum_{b=1}^{B} \sum_{i=1}^{N_b} \left(Y_{b_i} - \bar{Y} \right)^2 \right]}} \tag{9.10}$$

$$\text{Slope} = \beta = \frac{\sum_{b=1}^{B} \sum_{i=1}^{N_b} \left(X_{b_i} - \bar{X} \right) \left(Y_{b_i} - \bar{Y} \right)}{\sum_{b=1}^{B} \sum_{i=1}^{N_b} \left(X_{b_i} - \bar{X} \right)^2} \tag{9.11}$$

Equation 9.1 computes the overall number of observations by summing the number of observations in each stratum. Equation 9.2 gives the deviation for each observation. Equation 9.3 gives the Mean Deviation, which is the average deviation over all observations. Equation 9.4 takes the absolute value of Mean Deviation to compute the Quantity component of Mean Absolute Deviation (MAD). Equation 9.5 gives the Allocation Across Strata Deviation, which measures how the means of the strata differ. The Allocation Across Strata Deviation is positive when at least one stratum has a positive mean and at least one other stratum has a negative mean; otherwise, the Allocation Across Strata Deviation is zero. Equation 9.6 gives the Allocation Within Stratum Deviation, which compares how the deviations of the observations differ within each stratum. The Allocation Within Stratum Deviation is positive when both a positive deviation and a negative deviation exist among the observations within at least one stratum; otherwise, the Allocation Within Stratum Deviation is zero. Equation 9.7 gives Mean Absolute Deviation, which is the sum of the three components: Quantity, Allocation Across Strata, and Allocation Within Stratum. Equations 9.8, 9.9, 9.10 and 9.11 define metrics from the previous chapter.

Results at the bottom of Fig. 9.1 show that Mean Deviation is −0.08 °C during the half-year interval, which indicates overall cooling. Mean Deviation is 0.37 °C during the 28-year interval, which indicates overall warming. MAD during the half-year interval is 3.70 °C, and MAD during the 28-year interval is 0.74 °C. These results illustrate how Mean Deviation and MAD facilitate interpretation because they have the same units as **X** and **Y**. Figure 9.1 gives also the components of MAD. The Quantity component is the smallest of the three components during the half-year interval. The Quantity component is the largest of the three components during the 28-year interval. Moreover, the Quantity component during the half-year is smaller than the Quantity component during the 28 years. The Allocation Across Strata component is the largest of the three components during the half year because most of the seasonal change derives from warming in the northern hemisphere and cooling in the southern hemisphere. The Allocation Across Strata component is the zero during the 28 years when both the northern and southern hemispheres experience an increase in mean temperature. The Allocation Within Stratum component derives from simultaneous warming in some pixels and cooling in other pixels that reside within each of the hemispheres.

The scatter plots in Fig. 9.1 are helpful in a variety of respects. Each plot shows the center point, meaning the mean **X** and the mean **Y**. The center point for the half-year interval is (18.17, 18.09), which implies the Mean Deviation is −0.08, meaning overall cooling. The center point for the 28-year interval is (18.17, 18.54), which implies the Mean Deviation is 0.37, meaning overall warming. The scatter plots report also Correlation and Slope. Correlation indicates the strength and sign of a linear relationship between the start temperature and the end temperature for this example. Correlation is approximately 0.9 during the half-year change, while Correlation is stronger during the 28-year change. Slope summarizes how the end temperature relates to an increment increase in start temperature. The half-year interval has a positive slope that is less than one. This slope of the least squares line is flatter than the $Y = X$ line because lower start temperatures have larger positive deviations during the half-year interval. The 28-year interval has a slope that is within 0.011 of one, as the least squares line appears nearly parallel to the $Y = X$ line. This indicates that the temperatures at February 1982 do not influence substantially how the pixels experience warming. The blue points cover the red points in the scatter plots, which is one reason why metrics are essential to complement the visual assessment of the scatter plots.

The metrics reveal information that is not visually obvious in the scatter plots, because the scatter plots have 36,210 points, many of which are on top of each other. Each metric measures one characteristic concerning the arrangement of the points relative to the $Y = X$ line in the scatter plot. Visual examination of the scatter plots gives insights into the reasons for the results of the metrics. For example, the scatter plots show outliers that influence the metrics. Correlation and Slope derive from squared deviations as opposed to absolute deviations, thus larger deviations have a proportionally greater influence than smaller deviations on Correlation and Slope. The plot for the half-year interval shows outliers that are positive deviations at smaller start temperatures, which cause Slope to be less than one. The scatter plots show some characteristics that the metrics do not measure. For example, the upper right corner of the scatter plot for the half-year interval shows that the points are closer to $Y = X$ for larger start temperatures, which indicates that the absolute changes are smaller for larger start temperatures. Those larger temperatures derive from observations near the equator, which makes sense because the equator does not experience changes during seasons as strongly as other latitudes. Both metrics and scatter plots are important because the two forms of presentation complement each other. However, neither the metrics nor the scatter plots show the arrangement of the deviations in geographical space.

The maps show spatial information that neither the metrics nor the scatter plots show. Specifically, the greatest warming during the half-year interval is near China, northeastern USA, northern Europe, and also in the Mediterranean, Caspian, and Black seas. The greatest warming during the 28-year interval is in the southern hemisphere, especially to the west of Australia and to the east of South America. This case study illustrates why it is helpful to examine the data and results in the form of metrics, scatter plots, and maps. Each form offers insights that the other forms either ignore or fail to reveal clearly.

9.2 Discussion Questions

1. What do the metrics reveal that the scatter plots and maps do not reveal clearly?
2. What do the scatter plots reveal that the metrics and maps do not reveal clearly?
3. What do the maps reveal that the metrics and scatter plots do not reveal clearly?
4. How do the coordinates of the center point in the scatter plot relate to the Mean Deviation?
5. Allocation Across Strata Deviation is positive under what conditions?
6. Allocation Within Stratum Deviation is positive under what conditions?
7. What are additional applications for which the methods of this chapter would be enlightening?

Reference

Pontius Jr, R. G., Thontteh, O., & Chen, H. (2008). Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics, 15*, 111–142. https://doi.org/10.1007/s10651-007-0043-y.

Chapter 10
Indices of Agreement

Abstract This chapter examines seven indices that measure the agreement between two variables that show a phenomenon in the same units on an interval scale. Each index defines agreement by comparing the Observed Disagreement to a Baseline Disagreement, where each index has a distinct definition of Baseline Disagreement. This chapter uses the examples from Chap. 8 to illustrate characteristics of the indices. Mean Deviation, Mean Absolute Deviation, Correlation, and Slope have clearer and more helpful interpretations than the indices of agreement.

Keywords Agreement · Correlation · Index · Legates-McCabe · Nash-Sutcliffe · Slope · Willmott

10.1 Text

Chapter 8 recommends four metrics that measure interpretable characteristics of the patterns in a scatter plot of **Y** versus **X**, where **X** and **Y** each show a phenomenon in the same units on an interval scale. Those four metrics are Mean Deviation (MD), Mean Absolute Deviation (MAD), Correlation, and Slope. MD and MAD measure distinct aspects of the deviations between **X** and **Y**. Correlation and Slope measure distinct aspects of the association between **X** and its deviations with **Y**. This chapter compares those four metrics to seven indices of agreement.

This book focuses on difference because difference is clear and important for both categorical and interval variables. Difference is the same concept as disagreement for categorical variables. An observation from **X** disagrees with the corresponding observation from **Y** when the **X** category does not match the **Y** category. **X** agrees with **Y** when the **X** category matches the **Y** category. The sum of the sizes of disagreement and agreement equals the size of the extent for categorical variables, thus the size of the disagreement cannot be larger than the size of the extent. For interval variables, if disagreement means deviation, then disagreement is a clear concept. For example, the deviation between 3 and 2 equals 1. However, agreement is a vague concept for interval variables. The agreement between 3 and 2 does not make any sense unless we have a definition of the agreement between two numbers. Any index of agreement for interval variables must define agreement, which is

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1_10

conceptually challenging for several reasons. One reason is that the deviations of the **Y** values from their corresponding **X** values have no bounds for interval variables, unlike for categorical variables.

This chapter's indices of agreement define agreement as a function of Observed Disagreement and Baseline Disagreement in the form of Eq. 10.1. Subtraction between two numbers is a clear concept that defines the deviation for each observation. Therefore, Observed Disagreement could be clear depending on how Observed Disagreement summarizes deviations. This chapter's indices of agreement define Observed Disagreement as either the sum of absolute deviations or the sum of squared deviations, neither of which can be negative. However, there are other ways to summarize deviations, such as MD, which can be negative.

$$\text{Agreement} = 1 - \frac{\text{Observed Disagreement}}{\text{Baseline Disagreement}} = \frac{\text{Baseline Disagreement} - \text{Observed Disagreement}}{\text{Baseline Disagreement}} \quad (10.1)$$

For any definitions of Observed Disagreement and Baseline Disagreement, Eq. 10.1 has six properties. First, Observed Disagreement and Baseline Disagreement must have the same units for subtraction in Eq. 10.1 to make sense and for Agreement to be a unitless index, as is the case with this chapter's seven indices. Second, if Baseline Disagreement is zero, then Agreement is undefined. Third, if Baseline Disagreement is positive infinity, then Agreement equals one for any positive Observed Disagreement. Fourth, Agreement equals one if and only if the Observed Disagreement equals zero while the Baseline Disagreement does not equal zero. Fifth, Agreement equals zero if and only if the Observed Disagreement equals the Baseline Disagreement while both are not equal to zero. Sixth, Baseline Disagreement must be the maximum possible Observed Disagreement for Agreement to be bounded between zero and one. These properties create conceptual challenges to define Baseline Disagreement in a manner that corresponds to a particular definition of Observed Disagreement. Appropriate definitions of Observed Disagreement and Baseline Disagreement depend on the purpose of the index. Each index in this chapter has a distinct definition of Baseline Disagreement.

The inclusion of 1 and the subtraction of the ratio in Eq. 10.1 contributes no information from the data. The subtraction from 1 converts the ratio of differences into the concept of agreement, which is an unnecessary distraction that hinders interpretation. Interpretation in terms of disagreement would be clearer because Observed Disagreement and Baseline Disagreement define Agreement. The ratio after the second equals sign in Eq. 10.1 shows that Agreement measures how much smaller the Observed Disagreement is than the Baseline Disagreement, expressed as a proportion of Baseline Disagreement. This interpretation as a proportion is helpful, although I have not seen this interpretation in the literature. This interpretation reveals the awkwardness of Agreement because the ratio after the second equals sign is an unconventional way of expressing change from a baseline. The conventional way would have Observed Disagreement minus Baseline Disagreement in the numerator, which would allow clear interpretation in terms of reduction from a baseline.

Table 10.1 Properties of metrics where u indicates undefined due to division by zero

Metric	Range	Unit	Symmetric	Considers association	All $D_i = 0$ implies result	Result implies all $D_i = 0$
Mean Deviation	$(-\infty,\infty)$	X&Y	No	No	0	None
Mean Absolute Deviation	$[0,\infty)$	X&Y	Yes	No	0	0
Root Mean Squared Deviation	$[0,\infty)$	X&Y	Yes	No	0	0
Pearson's correlation r	$[-1,1]$	None	Yes	Yes	1 or u	None
Slope of least squares line β	$(-\infty,\infty)$	None	No	Yes	1 or u	None
Nash-Sutcliffe's E	$(-\infty,1]$	None	No	No	1 or u	1
Legates-McCabe's $E1$	$(-\infty,1]$	None	No	No	1 or u	1
Willmott's dr	$[-1,1]$	None	No	No	1 or u	1
Watterson's M	$[-1,1]$	None	Yes	Yes	1 or u	1
Mielke-Berry's $\mathfrak{R}$	$[-1,1]$	None	Yes	Yes	1 or u	1
Robinson's A	$[0,1]$	None	Yes	Yes	1 or u	1
Ji-Gallo's AC	$(-\infty,1]$	None	Yes	Yes	1 or u	1

Table 10.1 gives some properties of the four metrics that Chap. 8 recommends along with seven indices of agreement. This chapter describes also Root Mean Square Deviation (RMSD) because RMSD is a popular and frequently misinterpreted metric. RMSD is also known as Root Mean Squared Error when **X** is the truth and **Y** is a diagnosis. The Range column in Table 10.1 gives each metric's lower and upper bounds. The unit of MD, MAD, and RMSD is identical to the unit of both **X** and **Y**. Symmetric means the metric gives the same result when comparing **X** to **Y** as when comparing **Y** to **X**, which means that switching the definitions of **Y** and **X** does not affect the metric's result. A metric considers association when the metric's result depends on how each deviation is paired with each X value. MD, MAD, and RMSD do not consider association because they are functions of only the deviations, meaning those three metrics ignore how each deviation is paired with each X value. If all deviations equal zero, then the metric gives the result in the second column from the far right in Table 10.1, where u indicates undefined due to division by zero. For example, if all deviations equal zero, then Correlation is either one or undefined. The column on the far right in Table 10.1 specifies whether the result from the metric implies that all deviations equal zero. For example, if MAD equals zero, then all deviations must be zero. Whereas MD does not have a result that guarantees all deviations are zero. Other literature gives more details concerning these metrics' properties (Duveiller et al. 2016). Table 10.2 gives the mathematical notation for the metrics in Table 10.1. Observed Disagreement is the sum of absolute deviations for $E1$, dr, and $\mathfrak{R}$. Observed Disagreement is the sum of squared deviations for the other four indices of agreement.

Table 10.2 Mathematical notation for the metrics

Notation	Meaning
D_i	Deviation for observation *i*
$\bar{D}$	Mean of deviations **X**
i	Index for observation where $i = 1, 2, \ldots N$
j	Index for observation where $j = 1, 2, \ldots N$
N	Number of observations
X_i	Value of **X** for observation *i*
$\bar{X}$	Mean of **X**
Y_i	Value of **Y** for observation *i*
Y_j	Value of **Y** for observation *j*
$\bar{Y}$	Mean of **Y**

$$\bar{X} = \sum_{i=1}^{N} X_i \,/\, N \tag{10.2}$$

$$\bar{Y} = \sum_{i=1}^{N} Y_i \,/\, N \tag{10.3}$$

$$\text{Variance in } \mathbf{X} = \sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2 / \, N \tag{10.4}$$

$$\text{Variance in } \mathbf{Y} = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2 / \, N \tag{10.5}$$

$$D_i = Y_i - X_i \tag{10.6}$$

$$\text{Mean Deviation} = \text{MD} = \bar{D} = \bar{Y} - \bar{X} = \sum_{i=1}^{N} D_i \,/\, N \tag{10.7}$$

$$\text{Mean Absolute Deviation} = \text{MAD} = \sum_{i=1}^{N} \left|D_i\right| / \, N \tag{10.8}$$

$$\text{Root Mean Squared Deviation} = \text{RMSD} = \sqrt{\sum_{i=1}^{N} D_i^{\,2} \,/\, N} = \text{MAD}\sqrt{\frac{N\sum_{i=1}^{N} D_i^2}{\left(\sum_{i=1}^{N} \left|D_i\right|\right)^2}} \tag{10.9}$$

$$\text{Correlation} = r = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\left[\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2\right]\left[\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2\right]}} \tag{10.10}$$

$$\text{Slope} = \beta = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2} \tag{10.11}$$

$$E = 1 - \frac{\sum_{i=1}^{N}\left(X_i - Y_i\right)^2}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2} = 1 - \frac{\sum_{i=1}^{N} D_i^2}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2} = 1 - \frac{\text{RMSD}^2}{\text{Variance in }\mathbf{X}} \tag{10.12}$$

$$E1 = 1 - \frac{\sum_{i=1}^{N}\left|X_i - Y_i\right|}{\sum_{i=1}^{N}\left|X_i - \bar{X}\right|} = 1 - \frac{\sum_{i=1}^{N}\left|D_i\right|}{\sum_{i=1}^{N}\left|X_i - \bar{X}\right|} \tag{10.13}$$

$$dr = \left\{ \begin{array}{l} 1 - \dfrac{\sum_{i=1}^{N}\left|D_i\right|}{2\sum_{i=1}^{N}\left|X_i - \bar{X}\right|} \text{ when } \sum\limits_{i=1}^{N}\left|D_i\right| \le 2\sum\limits_{i=1}^{N}\left|X_i - \bar{X}\right| \\ \dfrac{2\sum_{i=1}^{N}\left|X_i - \bar{X}\right|}{\sum_{i=1}^{N}\left|D_i\right|} - 1 \text{ when } \sum\limits_{i=1}^{N}\left|D_i\right| > 2\sum\limits_{i=1}^{N}\left|X_i - \bar{X}\right| \end{array} \right\} \tag{10.14}$$

$$M = \left(\frac{2}{\pi}\right)\text{ARCSIN}\left[1 - \frac{\sum_{i=1}^{N} D_i^2}{\sum_{i=1}^{N}\left[\left(X_i - \bar{X}\right)^2 + \left(Y_i - \bar{Y}\right)^2 + \bar{D}^2\right]}\right] \tag{10.15}$$

$$\mathfrak{R} = 1 - \frac{N\sum_{i=1}^{N}\left|Y_i - X_i\right|}{\sum_{j=1}^{N}\sum_{i=1}^{N}\left|Y_j - X_i\right|} = 1 - \frac{\sum_{i=1}^{N}\left|D_i\right|}{\sum_{j=1}^{N}\sum_{i=1}^{N}\left|Y_j - X_i\right| / N} \tag{10.16}$$

$$A = 1 - \frac{\sum_{i=1}^{N} D_i^2}{\sum_{i=1}^{N}\left[\left(2X_i - \bar{X} - \bar{Y}\right)^2 + \left(2Y_i - \bar{X} - \bar{Y}\right)^2\right] / 2} \tag{10.17}$$

$$AC = 1 - \frac{\sum_{i=1}^{N} D_i^2}{\sum_{i=1}^{N}\left[\left(\left|\bar{D}\right| + \left|X_i - \bar{X}\right|\right)\left(\left|\bar{D}\right| + \left|Y_i - \bar{Y}\right|\right)\right]} \tag{10.18}$$

Equation 10.12 is Nash-Sutcliffe's *E*, which is popular in hydrology but not specific to hydrology (Jackson et al. 2019; Nash and Sutcliffe 1970). If all the **X** values are identical, then the *E* is undefined. If the **X** values are not identical and all the deviations are zero, then Nash-Sutcliffe's *E* is one. If Nash-Sutcliffe's *E* equals one, then all deviations are zero. Nash-Sutcliffe's *E* has no lower bound when the deviations between **Y** and **X** have no bound. Nash-Sutcliffe's *E* is not symmetric, as **X** appears in the ratio's denominator but **Y** does not. Nash-Sutcliffe's *E* does not consider the association between the deviations and the **X** values because the Observed Disagreement derives from only the deviations and the Baseline Disagreement derives from only **X**. Nash-Sutcliffe's *E* is one minus the ratio of the square of RMSD to the Variance in **X**.

Equation 10.13 is Legates-McCabe's *E1* (Legates and McCabe 1999). *E1* uses absolute deviations where Nash-Sutcliffe's *E* uses squared deviations. Thus *E1* has some of the same properties as *E*. Specifically, Legates-McCabe's *E1* is not symmetric and does not consider the association between the deviations and the **X** values.

Equation 10.14 is Willmott's *dr*, which is the most recent in a sequence of indices that Willmott developed over decades (Willmott et al. 2012). Willmott's *dr* is conceptually similar to Legates-McCabe's *E1*. However, *E1* has no lower bound, whereas −1 is the lower bound of *dr* because *dr* considers two cases. Willmott's *dr* is not symmetric and does not consider the association between the deviations and the **X** values. Authors debate the merits of *dr* relative to *E1* and the other indices of agreement (Jackson et al. 2019; Legates and McCabe 2013).

Equation 10.15 is Watterson's *M* (Watterson 1996). Waterson's *M* uses the ARCSIN function to transform an expression that has the form of Eq. 10.1. ARCSIN is a non-linear increasing function for which the domain is [−1,1] and the range is [−π/2,π/2]. Multiplication by (2/π) in the equation for Watterson's *M* makes the range for *M* become [−1,1]. *M* is symmetric, as are the remaining indices in Table 10.1. Watterson confirmed via personal communication that equation 28 of Watterson (1996) has a typographical error while Eq. 10.8 of Duveiller et al. (2016) gives the correct equation for *M*, which is equivalent to this book's Eq. 10.15.

Equation 10.16 is a version of Mielke-Berry's R that uses absolute deviations (Mielke Jr et al. 1996; Mielke Jr and Berry 2007). Berry and Mielke refer to $\mathfrak{R}$ as "curly r". The denominator sums the absolute deviations for all possible ways to pair each **X** value with each **Y** value, while the numerator sums the absolute deviations for only the observed pairings between each **X** value with each **Y** value. If the observed pairings were random, then the expected numerator would equal the denominator, which would cause $\mathfrak{R}$ to be zero. If the data were to pair relatively large values in **X** with relatively large values in **Y**, then the numerator would be smaller than the denominator, which would cause $\mathfrak{R}$ to be positive.

Equation 10.17 is equivalent to Robinson's *A* (Robinson 1957). The range of *A* is [0,1] and *A* is symmetric. If *A* equals one, then all D_i are zero.

Equation 10.18 is Ji-Gallo's *AC* (Ji and Gallo 2006). The equation's format shows that *AC* treats **Y** in the same manner as *AC* treats **X**, thus *AC* is symmetric. If *AC* equals one, then all D_i are zero.

The top of Fig. 10.1 shows the example data from Chap. 8 for nine series of **Y**. MD equals 0 for series A-D. MD equals −3 for series E-F. MD equals −4 for series G-I. MAD equals 4 for all series. A visual inspection of the scatter plots might lead readers to consider a particular series to have more agreement with **X** than other series, depending on the reader's intuition concerning the meaning of agreement. Parts a-i in Fig. 10.1 plot the results for Correlation, Slope, and seven indices versus MD. Each letter A-I in each plot denotes one of the nine series of **Y** in the four scatter plots at the top of the figure. Correlation for series A and C are so similar that the letters overlap in Fig. 10.1a. Similarly, Correlation for series B and D are so similar that the letters overlap in Fig. 10.1a. Correlation equals one for only series I. Slope is positive for series A, C, E, G, and I, while Slope is negative for the other series. Slope equals one for only series I. Nash-Sutcliffe's *E* equals −5.4 for series A, B, and I, which have RMSD equal to 4; while Nash-Sutcliffe's *E* equals −9.0 for the other series, which have RMSD equal to 5. For all series, Legates-McCabe's *E1*

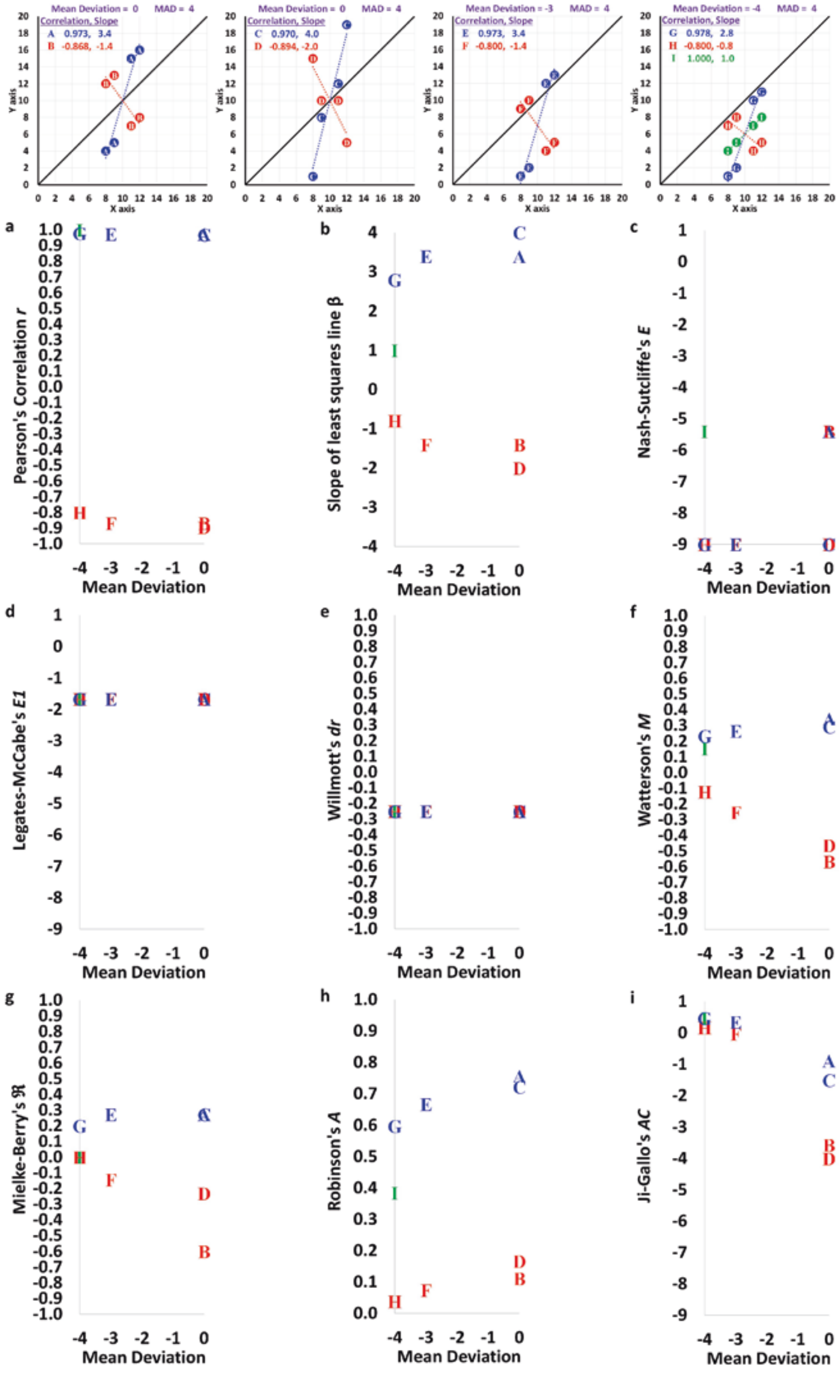


Fig. 10.1 Results for metrics where each letter A-I in each plot denotes a series for **Y**

equals −1.67 while Willmott's *dr* equals −0.25. Legates-McCabe's *E1* and Willmott's *dr* are constant across series because all series have the same MAD and the same **X**. The pattern for Waterson's *M* is similar to the pattern for Slope; specifically, the sign of *M* is the same as the sign of Slope for each series. Mielke-Berry's $\mathfrak{R}$ equals zero for series H and I. Series A, C, and E have the same $\mathfrak{R}$. Robinson's *A* gives larger values when Slope is positive than when Slope is negative. Robinson's *A* increases as MD becomes closer to zero. Ji-Gallo's *AC* is negative for four of the nine series, even though their publication claims that *AC* is bounded between 0 and 1 (Ji and Gallo 2006). As MD becomes closer to zero, *AC* tends to decrease. The only difference between series A & B is how the deviations are associated with the **X** values. The same is true for the difference between series C & D, and E & F. Figure 10.1 shows that *M*, $\mathfrak{R}$, *A*, and *AC* differentiate between these pairs of series, which illustrates how those indices consider association. I list below seven concepts that derive from comparing the characteristics of the indices of agreement to characteristics of the four recommended metrics: MD, MAD, Correlation, and Slope.

First, the concept of a difference is clear for categorical and interval variables, while the concept of agreement is vague for interval variables. This is the reason why the title of this book and most of its contents concern difference rather than agreement. Disagreement has a clear interpretation for interval variables as Eq. 10.6 shows. Equation 10.1 for agreement is bizarre in the respect that it expresses agreement in terms of two types of disagreement: Observed Disagreement and Baseline Disagreement. Equation 10.1 includes Baseline Disagreement so that the index of agreement equals zero when Observed Disagreement equals Baseline Disagreement, assuming Baseline Disagreement is not equal to zero. Each index of agreement has a distinct definition of Baseline Disagreement. Some definitions are conceptually complex, thus mathematically complicated and challenging to interpret. It is not necessarily clear what Baseline Disagreement would be appropriate for a particular application, or whether a Baseline Disagreement is relevant.

Second, if a metric measures exactly one characteristic of the pattern in a scatter plot, then the metric has a clear interpretation. MD measures one characteristic, which is average **Y** minus average **X**. MAD is the average vertical distance from the $Y = X$ line to the points in a scatter plot. Correlation gives the strength and sign of a linear relationship between **X** and **Y**. Slope is the change in **Y** for each increment increase in **X** for the least squares line. MD, MAD, Correlation, and Slope each measure one characteristic of a scatter plot's pattern. Figure 10.1 reveals how *E1* and *dr* show no variation among cases A-I because *E1* and *dr* define Observed Disagreement in terms of MAD, which is identical for cases A-I. If a metric integrates more than one characteristic of the pattern, then the metric's interpretation is challenging. Graphs in Fig. 10.1f–i show how those indices integrate various characteristics of pattern, which hinders the interpretation of *M*, $\mathfrak{R}$, *A*, and *AC*. If any of those indices are less than 1, then it is not immediately clear what characteristic of the pattern is responsible.

Third, the sum of squared deviations is the Observed Disagreement in four of the indices: E, M, A, and AC. The sum of squared deviations is equal to N times the square of RMSD. However, RMSD is difficult to interpret because RMSD lacks a straightforward interpretation in terms of the distances in the scatter plot because RMSD is a function of the sum of squared deviations, not absolute deviations. This book includes RMSD as a warning to readers who might see RMSD in other literature. Scientists frequently mistakenly interpret RMSD as if RMSD were the average distance between the $Y = X$ line and the points in the scatter plot, which RMSD is not, but MAD is. MAD = RMSD if and only if the absolute values of all D_i are identical. Otherwise, MAD < RMSD. If MAD = RMSD, then this does not necessarily imply that all D_i are identical, as series A and B illustrate in Fig. 10.1. RMSD combines MAD with the variation among the absolute values of D_i into one metric, which hinders interpretation of RMSD for practical applications (Pontius Jr et al. 2008; Willmott et al. 2009; Willmott and Matsuura 2005, 2006). Equation 10.9 shows that RMSD equals MAD times the square root of a ratio when the denominator of the ratio is not zero. If the absolute value of all the deviations are identical, then the numerator equals the denominator; otherwise, the numerator is greater than the denominator. RMSD is difficult to interpret because RMSD uses squared deviations, just as four of the seven indices use squared deviations to compute Observed Disagreement. Thus, those four indices suffer from the same difficulty of interpretation that RMSD does. Several scientists have told me that a motivation to square a deviation is to convert negative deviations to positive deviations, so negative deviations do not cancel with positive deviations during the sum of deviations. If the goal is to convert negative deviations into positive deviations, then the absolute value is a simpler way than squaring to accomplish that goal. Other scientists have told me that a motivation to square the deviations is to place a disproportionally larger weight on larger absolute deviations when computing Observed Disagreement. However, it is not clear why that would be a desirable characteristic, or why squaring would be any more desirable than the infinite number of other ways to place disproportionally larger weights on larger absolute deviations.

Fourth, we must understand an equation to interpret its result; simpler equations are easier to understand. The equations for MD, MAD, Correlation, and Slope are simpler than the equations for most of this chapter's indices of agreement. Consequently, MD, MAD, Correlation, and Slope have clearer interpretations than most of the indices. Users should know the mathematical properties of an equation before attempting to interpret the result of the equation. Complex equations can be confusing to even their inventors. For example, Ji and Gallo (2006) claim that the range of their index is [0,1], but series A-D produce values that are outside that range.

Fifth, the apparent motivation of some indices is to compare a novel method of prediction to a baseline prediction that derives from a previously established baseline method of prediction. However, an index to compare a novel method of prediction **Y** to a baseline method of prediction **B** should use the truth **X** and the prediction **Y** to generate the Observed Disagreement then use **X** and **B** to compute the Baseline Disagreement. But this chapter's indices of agreement use **X** and **Y** to generate both the Observed Disagreement and the Baseline Disagreement, so none of the indices are suited to compare a novel method of prediction to a baseline method of prediction. For example, Eqs. 10.12 for *E* and 10.13 for *E1* show the average **X** in the ratio's denominator plays the role of the values for **Y** in the ratio's numerator. In those cases, the average **X** is the baseline prediction, which violates the definition of prediction. A prediction by definition exists before the truth is known. The truth must not be used to generate a prediction. The same problem exists for the other indices of agreement because all of the indices use **X** to compute the Baseline Disagreement. Consequently, it is not clear how to interpret any of this chapter's indices for applications that relate to predictive accuracy.

Sixth, one must consider whether the index relates to the particular goal or research question. The authors who have derived the indices of agreement in this chapter have dedicated substantial effort to derive their indices with particular goals or research questions in mind. A scientist must consider whether a particular index gives information relevant to the research question. Let us consider the properties of Nash-Sutcliffe's *E*, which hydrologists use to compare data to outputs from simulation models. Equation 10.12 shows that *E* is a function of RMSD and the variance in **X**. If several simulations for **Y** are compared to the same **X** values, then those comparisons have the same variance in **X**. Therefore, only RMSD determines how *E* would distinguish among the cases. If we interpreted only RMSD, then we could interpret the results in the same units as **X** and **Y**. If the purpose is to compare several simulations for the same **X**, then RMSD offers a clearer comparison than *E*. *E* might be helpful when one wishes to compare a case study for one **X** to a case study for a different **X**. But for that type of comparison, *E* would be a relevant index when the only criterion is the size of the square of RMSD relative to the size of the variance in **X**. *E* ignores many other patterns, such as the how the deviations are paired with the **X** values, which Correlation and Slope measure. *E* might be relevant when the association between the deviations and the **X** values are irrelevant to the research question. But if I were to evaluate a simulation model's output, then my first question would likely concern Mean Deviation, which *E* does not necessarily indicate. I have asked hydrologists why they report *E*. They tell me that *E* is part of the culture among hydrologists. Their response is a social reason, not a scientific reason.

Seventh, some of the indices use a form of randomness to generate the Baseline Disagreement. In this case, the index would be relevant for situations where the particular form of randomness is relevant to the research question. If randomness is not important to the research question, then the scientist should not use a metric that has a baseline of randomness. A baseline of randomness is a distraction in most of the practical applications that I have seen. For example, if the purpose of the index is to measure the accuracy of a novel method of diagnosis, then the Baseline Disagreement should derive from a previously established method of diagnosis. Established methods do not diagnose randomly.

I can understand an initial desire to use an index of agreement. I have spent decades researching and deriving indices of agreement, mostly for categorical variables. Two of my most highly cited publications proposed three indices of agreement (Pontius Jr 2000, 2002). The indices of agreement related to Kappa, which is a popular index that has the form of Eq. 10.1 where the Baseline Disagreement derives from a form of randomness. Over time, I have found that the indices of agreement have been horrendously misleading because the indices compared Observed Disagreement to an irrelevant baseline. I regretted so deeply publishing those indices of agreement that I apologized a decade later and asked colleagues to abandon them (Pontius Jr and Millones 2011). One of my equations was so complex that even I did not fully appreciate its mathematical properties until I saw how other scientists applied it. I made conceptual blunders and I learned from my blunders. I write this book to spare others the waste of making the same blunders. However, I have found some authors reluctant to abandon kappa, even when their articles cite my paper that has the title "Death to Kappa" (Pontius Jr and Millones 2011). My discussions with authors reveal that authors continue to use kappa because of their mindless habits and their assumption that readers have the same habits. The continued reporting of kappa reinforces the existing poor practice and indoctrinates the next generation of scientists in the same dysfunctional thinking. Science does not progress in such a culture.

My main mistake earlier in my career was to focus on agreement, rather than difference. My focus on agreement inspired me to define various baselines, for which I used various forms of randomness. Comparison to randomness was a theme in my formal statistical education, so I thought I was deriving helpful indices. Comparison to randomness might be important for a limited scope of research questions. For example, if a casino is trying to find cheaters in games where the players attempt to select random numbers, then the casino would search for players who win significantly more than random expectation. Another example might be when an investigator wants to measure how a machine learning algorithm diagnoses randomly selected testing data. However, comparison to randomness is irrelevant for many practical research questions. I have seen many practical applications where the results have Observed Disagreements that are important regardless of their differences from a random baseline, in which case the comparison to randomness is a distraction from the important results. Observed Disagreements between a diagnosis and the truth reveal diagnostic errors, which reveal opportunities to improve the method of diagnosis. Comparison to randomness might be relevant when the alternative is a random diagnosis. I have not seen cases in my career as an applied statistician where the alternative procedure is to diagnose randomly. A diagnosis can deviate substantially from random but deviate in important respects from perfect. It is frequently more important to know how a diagnosis deviates from perfection than how a diagnosis deviates from random. If the scientist already knows that the phenomenon that **X** describes is not random, then a comparison of **Y** to a random baseline is a distraction.

I have not seen cases in my professional experience where indices of agreement are more useful than the collection of MD, MAD, Correlation, and Slope. Moreover, I have reviewed journal articles in which indices of agreement failed to reveal important patterns that simpler metrics would have revealed. I have seen authors routinely give wrong or unhelpful interpretations of indices of agreement. A major challenge with indices of agreement is that they require a Baseline Disagreement for comparison to the Observed Disagreement. But it is frequently not clear how to establish an appropriate Baseline Disagreement. Many applications lack a natural baseline and do not require a baseline. For example, the analysis of temporal change does not require a baseline because Observed Disagreement indicates temporal change, which is straightforward without a baseline. MD, MAD, Correlation, and Slope do not require a baseline, thus avoid many of the conceptual problems with indices of agreement.

10.2 Discussion Questions

1. The difference between 2 and 5 equals what?
2. The agreement between 2 and 5 equals what?
3. What is a fundamental difference between disagreement for a categorical variable versus disagreement for an interval variable?
4. How does Eq. 10.1 compare to the conventional mathematical expression to express change from a baseline?
5. What are the challenges in constructing an index of agreement?
6. How does Fig. 10.1 reveal the metrics that consider the association between the **X** values and the deviations as opposed to the indices that do not consider the association?
7. If all D_i are equal, then does Mean Absolute Deviation equal Root Mean Squared Deviation? If not, give a counterexample.
8. If Mean Absolute Deviation equals Root Mean Squared Deviation, then are all D_i equal? If not, give a counterexample.
9. Under what conditions would comparison to a random baseline be a distraction?
10. What would motivate a scientist to use one of the indices of agreement in this chapter?

References

Duveiller, G., Fasbender, D., & Meroni, M. (2016). Revisiting the concept of a symmetric index of agreement for continuous datasets. *Scientific Reports, 6*, 19401. https://doi.org/10.1038/srep19401.

Jackson, E. K., Roberts, W., Nelsen, B., Williams, G. P., Nelson, E. J., & Ames, D. P. (2019). Introductory overview: Error metrics for hydrologic modelling – A review of common practices and an open source library to facilitate use and adoption. *Environmental Modelling & Software, 119*, 32–48. https://doi.org/10.1016/j.envsoft.2019.05.001.

Ji, L., & Gallo, K. (2006). An agreement coefficient for image comparison. *Photogrammetric Engineering and Remote Sensing, 72*, 823–833. https://doi.org/10.14358/PERS.72.7.823.

Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research, 35*, 233–241. https://doi.org/10.1029/1998WR900018.

Legates, D. R., & McCabe, G. J. (2013). A refined index of model performance: A rejoinder. *International Journal of Climatology, 33*, 1053–1056. https://doi.org/10.1002/joc.3487.

Mielke, P. W., Jr., & Berry, K. J. (2007). *Permutation methods: A distance function approach* (Springer series in statistics) (2nd ed.). New York: Springer.

Mielke, P. W., Jr., Berry, K. J., Landsea, C. W., & Gray, W. M. (1996). Artificial skill and validation in meteorological forecasting. *Weather and Forecasting, 11*, 17. https://doi.org/10.1175/1520-0434(1996)011<0153:ASAVIM>2.0.CO;2.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I – A discussion of principles. *Journal of Hydrology, 10*, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Pontius Jr, R. G. (2000). Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing, 66*, 1011–1016.

Pontius Jr, R. G. (2002). Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. *Photogrammetric Engineering and Remote Sensing, 68*, 1041–1050.

Pontius Jr, R. G., & Millones, M. (2011). Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing, 32*, 4407–4429. https://doi.org/10.1080/01431161.2011.552923.

Pontius Jr, R. G., Thontteh, O., & Chen, H. (2008). Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics, 15*, 111–142. https://doi.org/10.1007/s10651-007-0043-y.

Robinson, W. S. (1957). The statistical measurement of agreement. *American Sociological Review, 22*, 17. https://doi.org/10.2307/2088760.

Watterson, I. G. (1996). Non-dimensional measures of climate model performance. *International Journal of Climatology, 16*, 379–391. https://doi.org/10.1002/(SICI)1097-0088(199604)16:4<379::AID-JOC18>3.0.CO;2-U.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*, 79–82. https://doi.org/10.3354/cr030079.

Willmott, C. J., & Matsuura, K. (2006). On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International Journal of Geographical Information Science, 20*, 89–102. https://doi.org/10.1080/13658810500286976.

Willmott, C. J., Matsuura, K., & Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment, 43*, 749–752. https://doi.org/10.1016/j.atmosenv.2008.10.005.

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology, 32*, 2088–2094. https://doi.org/10.1002/joc.2419.

Chapter 11
Vector Variable Versus Vector Variable

Abstract This chapter shows how to compute components of deviation when **X** and **Y** are vectors meaning each observation has magnitude and direction. This chapter shows how to compute Mean Deviation and Mean Absolute Deviation for magnitude and direction. Vector addition helps to compute the mean **X** and the mean **Y**. An example illustrates the concepts. Relevant software includes Vector Difference Analysis in Google Earth Engine via https://taoshiqi.users.earthengine.app/view/wind-vector-comparison (Tao and Bajracharya 2020) and VectorDeviation in R available at https://github.com/skievanc/VectorDeviation (Collins, 2020).

Keywords Direction · Magnitude · Mean Absolute Deviation · Vector variable

11.1 Text

This chapter uses the concepts of Chap. 8 to give methods to compare variables that are vectors. A vector is a type of variable that has magnitude, where positive magnitudes have also direction. Applications include phenomena that have both magnitude and direction, such as wind (Peng et al. 2013). Another application is to topography where each vector's magnitude is steepness and its direction is aspect. Most of the applications that I envision concern geographical phenomena; therefore, this chapter illustrates direction as the number of degrees increasing clockwise from north, as opposed to the number of radians increasing counterclockwise from a horizontal axis. For example, if we were to study ocean currents, then the magnitude would be the current's speed while the direction would be degrees from north. We might want to know how a predicted current compares to the observed current at various places, or we might want to know how the current changes from the start time to the end time.

Figure 11.1 illustrates the concepts. The example data consist of four pixels, where each pixel contains a vector for **X** and a vector for **Y**. The number next to each vector denotes the vector's magnitude, which is proportional to the arrow's length. The dashed vectors show the north and east components. For example, the northwestern pixel for **X** has a vector with a magnitude of 30 times the square root of 2 and a direction of 45 degrees from north. The vector's north component is 30 and its east component is 30. The northeastern pixel for **X** has a vector with a

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1_11

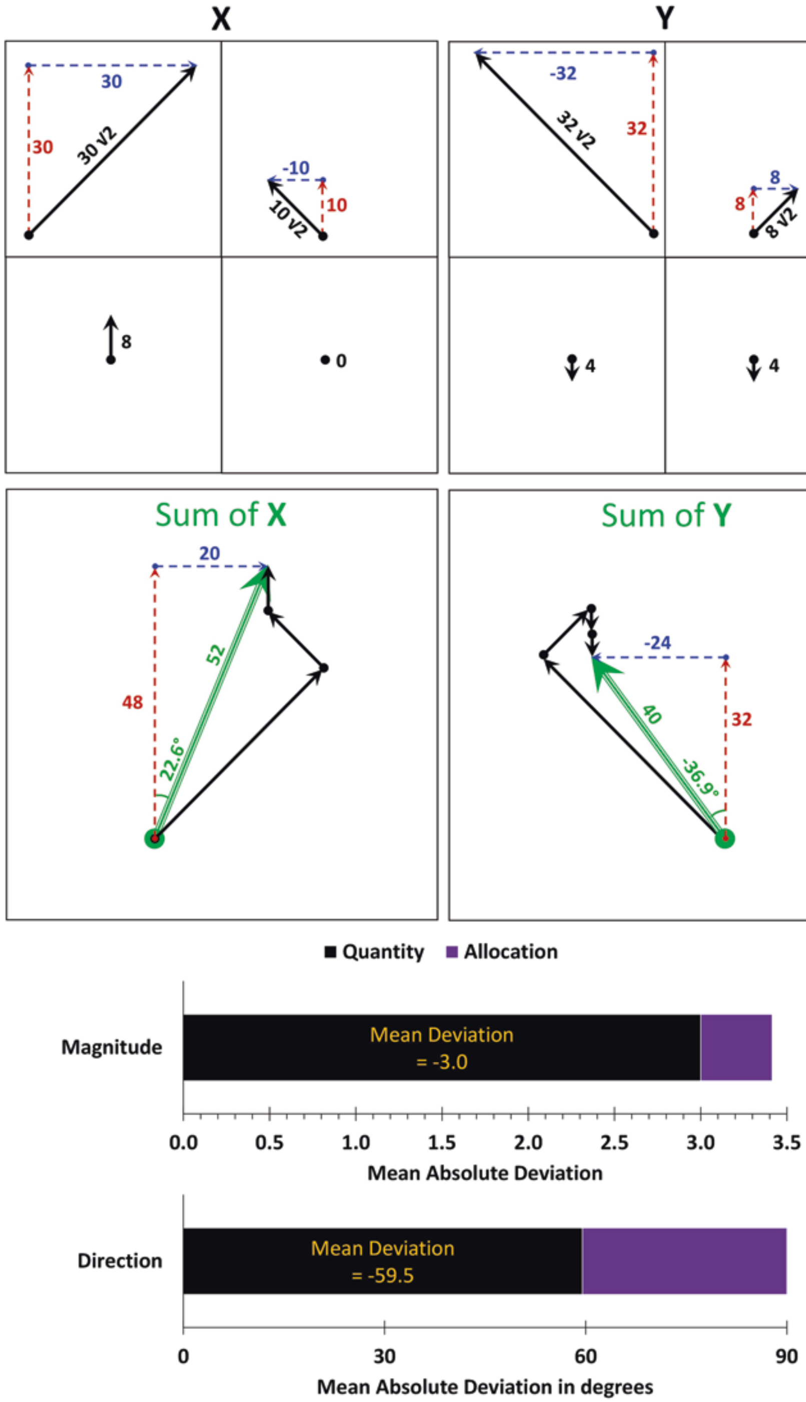


Fig. 11.1 Example to compare **X** and **Y** that show a vector phenomenon

magnitude of 10 times the square root of 2 and a direction of −45 degrees from north. The southwestern pixel for **X** has a vector with a magnitude of 8 and a direction of 0 degrees from north. The southeastern pixel for **X** has the zero vector, which has 0 magnitude and no direction. The northwestern pixel for **Y** has a vector with a magnitude of 32 times the square root of two and a direction −45 degrees from north. Thus the deviation in magnitude between **X** and **Y** in the northwestern pixel is 2 times the square root of 2 in terms of magnitude and − 90 degrees in terms of direction. The deviation in direction is negative because the vector in **Y** is counterclockwise from the corresponding vector in **X**. The deviation in magnitude between **X** and **Y** is −2 times the square root of two in the northwestern pixel, −4 in the southwestern pixel, and 4 in the southeastern pixel. The deviation in direction between **X** and **Y** is 90 degrees in the northeastern pixel, 180 degrees in the southwestern pixel, and 0 degrees in the southeastern pixel.

We take an approach similar to Chap. 8 where we used the means of **X** and **Y** to compute the Mean Deviation and components of Mean Absolute Deviation (Pontius et al. 2008). The mean of a real variable is the variable's values summed over all observations divided by the number of observations. The mean of a vector variable uses vector addition to compute the sum of vectors. Vector addition sums vectors by stringing together the sequence of vectors where the tip of each vector connects to the start of another vector. The sum is the resultant vector drawn from the start of the first vector to the tip of the last vector in the string's sequence. The sum has a magnitude and direction. For example, the sum of **X** in Fig. 11.1 has a magnitude of 52 consisting of a north component of 48 and an east component of 20. Four observations produce the sum that has magnitude 52, thus the mean magnitude is 52 divided by 4, which is 13. The mean of **X** has a direction of ARCTAN(20/48), which is approximately 22.6 degrees from north. The mean of **Y** in Fig. 11.1 has magnitude 10 and direction ARCTAN(−24/32), which is approximately −36.9 degrees from north. Table 11.1 gives the mathematical notation that the equations use.

Equations 11.1, 11.2 and 11.3 define the mean magnitude for the vectors in **X**. Equation 11.1 sums the north components of the individual vectors. Equation 11.2 sums the east components of the individual vectors. Equation 11.3 uses the Pythagorean Theorem to compute the mean magnitude of the mean vector. Equation 11.4 uses the ARCTAN function to compute the direction of the mean vector. The ARCTAN function gives the number of degrees of an angle in a right triangle as a function of the ratio of the length of the triangle's side opposite the angle to the length of the triangle's side adjacent to the angle. The ARCTAN function has a domain of (−∞,∞) and range of (−90°,90°). The first case in Eq. 11.4 applies when the mean vector has a positive north component, thus gives degrees in the range of (−90°,90°). The second case in Eq. 11.4 gives an angle's degrees in the range [90°,180°) because the ARCTAN function for the second case gives a result in the range (−90°,0°]. The third case gives an angle's number of degrees in the range [−90°,−180°) because the ARCTAN function for the third case gives a result in the range [0°,90°). The fourth case gives 180° when the mean vector points directly south. The fifth case is when the magnitude of the mean vector is zero, in which case

Table 11.1 Notation to compare two variables that show the same vector phenomenon

Notation	Possible values	Meaning
i	1, 2, …, N	Index of vector
N	Positive integer	Number of vectors
X_i	$[0,\infty)$	Magnitude of vector i in **X**
Xe	$[0,\infty)$	Sum of east magnitudes of vectors in **X**
Xn	$[0,\infty)$	Sum of north magnitudes of vectors in **X**
$\overline{X}$	$[0,\infty)$	Magnitude of the mean vector for **X**
Y_i	$[0,\infty)$	Magnitude of vector i in **Y**
Ye	$[0,\infty)$	Sum of east magnitudes of vectors in **Y**
Yn	$[0,\infty)$	Sum of north magnitudes of vectors in **Y**
$\overline{Y}$	$[0,\infty)$	Magnitude of the mean vector for **Y**
θ_i	$(-180°,180°]$	Direction of vector i in **X**. If $X_i = 0$, then θ_i does not exist.
$\overline{\theta}$	$(-180°,180°]$	Direction of the mean vector for **X**
φ_i	$(-180°,180°]$	Direction of vector i in **Y**. If $Y_i = 0$, then φ_i does not exist.
$\overline{\varphi}$	$(-180°,180°]$	Direction of the mean vector for **Y**
$\lvert\delta_i\rvert$	$[0°,180°]$	Absolute direction deviation for vector pair i
$\overline{\delta}$	$(-180°,180°]$	Direction deviation between mean vectors

the mean vector does not have a direction. Equations 11.5, 11.6, 11.7 and 11.8 compute for **Y** the concepts that Eqs. 11.1, 11.2, 11.3 and 11.4 compute for **X**. Equation 11.9 computes the magnitude of the deviation between the mean vectors for **X** and **Y**. If the magnitude of the mean vector for **Y** is smaller than the magnitude of the mean vector for **X**, then the magnitude of their deviation is negative. Equation 11.10 computes the Quantity component of the magnitude of the Mean Absolute Deviation (MAD) following the logic of Chap. 8. Equation 11.11 computes the magnitude's Allocation component of MAD for all vector pairs. Equation 11.11 uses the MAXIMUM function to assure the Allocation component is not negative. The properties of vector addition can cause the second argument in the MAXIMUM function to be negative. Equation 11.12 gives the magnitude's MAD such that it is the sum of its components of Quantity and Allocation, conceptually similar to how Chap. 8 defines MAD as the sum of its components. Equation 11.12 uses the MAXIMUM function for the same reason Eq. 11.11 uses the MAXIMUM function.

$$Xn = \sum_{i=1}^{N}\left[X_i \mathrm{COS}\left(\theta_i\right)\right] \tag{11.1}$$

$$Xe = \sum_{i=1}^{N}\left[X_i \mathrm{SIN}\left(\theta_i\right)\right] \tag{11.2}$$

$$\overline{X} = \left(\sqrt{Xn^2 + Xe^2}\right) / N \tag{11.3}$$

$$\overline{\theta} = \left\{ \begin{array}{c} \text{ARCTAN}\left(Xe / Xn\right) \text{when } Xn > 0 \\ 90^{\circ} - \text{ARCTAN}\left(Xn / Xe\right) \text{when } Xn \leq 0 \text{ and } Xe > 0 \\ -90^{\circ} - \text{ARCTAN}\left(Xn / Xe\right) \text{when } Xn \leq 0 \text{ and } Xe < 0 \\ 180^{\circ} \text{ when } Xn < 0 \text{ and } Xe = 0 \\ \text{does not exist when } Xn = 0 \text{ and } Xe = 0 \end{array} \right\} \tag{11.4}$$

$$Yn = \sum_{i=1}^{N} \left[Y_i \text{COS}\left(\varphi_i\right) \right] \tag{11.5}$$

$$Ye = \sum_{i=1}^{N} \left[Y_i \text{SIN}\left(\varphi_i\right) \right] \tag{11.6}$$

$$\overline{Y} = \left(\sqrt{Yn^2 + Ye^2} \right) / N \tag{11.7}$$

$$\overline{\varphi} = \left\{ \begin{array}{c} \text{ARCTAN}\left(Ye / Yn\right) \text{when } Yn > 0 \\ 90^{\circ} - \text{ARCTAN}\left(Yn / Ye\right) \text{when } Yn \leq 0 \text{ and } Ye > 0 \\ -90^{\circ} - \text{ARCTAN}\left(Yn / Ye\right) \text{when } Yn \leq 0 \text{ and } Ye < 0 \\ 180^{\circ} \text{ when } Yn < 0 \text{ and } Ye = 0 \\ \text{does not exist when } Yn = 0 \text{ and } Ye = 0 \end{array} \right\} \tag{11.8}$$

$$\text{Magnitude Mean Deviation} = \overline{Y} - \overline{X} \tag{11.9}$$

$$\text{Magnitude Quantity Deviation} = \left| \overline{Y} - \overline{X} \right| \tag{11.10}$$

$$\text{Magnitude Allocation Deviation} = \text{MAXIMUM}\left[0, \left(\sum_{i=1}^{N} \left| Y_i - X_i \right| / N \right) - \left| \overline{Y} - \overline{X} \right| \right] \tag{11.11}$$

$$\text{Magnitude Mean Absolute Deviation} = \text{MAXIMUM}\left(\left| \overline{Y} - \overline{X} \right|, \sum_{i=1}^{N} \left| Y_i - X_i \right| / N \right) \tag{11.12}$$

$$\left| \delta_i \right| = \left\{ \begin{array}{c} \text{MINIMUM}\left(\left| \varphi_i - \theta_i \right|, 360^{\circ} - \left| \varphi_i - \theta_i \right| \right) \text{when } X_i > 0 \text{ and } Y_i > 0 \\ 0^{\circ} \text{ when } X_i = 0 \text{ or } Y_i = 0 \end{array} \right\} \tag{11.13}$$

$$\overline{\delta} = \left\{ \begin{array}{c} \overline{\varphi} - \overline{\theta} \text{ when } \left| \overline{\varphi} - \overline{\theta} \right| < 180^{\circ} \\ \overline{\varphi} - \overline{\theta} - 360^{\circ} \text{ when } \overline{\varphi} - \overline{\theta} > 180^{\circ} \\ \overline{\varphi} - \overline{\theta} + 360^{\circ} \text{ when } \overline{\varphi} - \overline{\theta} < -180^{\circ} \\ 180^{\circ} \text{ when } \left| \overline{\varphi} - \overline{\theta} \right| = 180^{\circ} \\ 0^{\circ} \text{ when } \overline{\varphi} \text{ or } \overline{\theta} \text{ do not exist} \end{array} \right\} \tag{11.14}$$

$$\text{Direction Quantity Deviation} = \left|\bar{\delta}\right| \tag{11.15}$$

$$\text{Direction Allocation Deviation} = \text{MAXIMUM}\left[0, \left(\sum_{i=1}^{N}\left|\delta_i\right|/N\right) - \left|\bar{\delta}\right|\right] \tag{11.16}$$

$$\text{Direction Mean Absolute Deviation} = \text{MAXIMUM}\left(\left|\bar{\delta}\right|, \sum_{i=1}^{N}\left|\delta_i\right|/N\right) \tag{11.17}$$

Equation 11.13 gives the absolute deviation between vector Y_i and vector X_i in terms of direction. The deviation in direction derives from an angle that vectors Y_i and X_i form. If the angle of Y_i minus the angle of X_i is in the interval [−180°,180°], then the absolute deviation is the first argument in the MINIMUM function of Eq. 11.13. If the angle of Y_i minus the angle of X_i is outside the interval [−180°,180°], then the absolute deviation is the second argument in the MINIMUM function of Eq. 11.13. Therefore, Eq. 11.13 produces an absolute deviation in the interval [0°,180°]. Equation 11.13 gives zero when either vector has zero magnitude.

Equation 11.14 computes the deviation in direction of the mean vectors for **X** and **Y**. The deviation derives from the angle that the pair of mean vectors form with absolute size on the interval [0°,180°]. If the angle's mean vector for **Y** is clockwise from the mean vector for **X**, then the direction's mean deviation is positive. If the angle's mean vector for **Y** is counterclockwise from the mean vector for **X**, then the direction's mean deviation is negative. The first case in Eq. 11.14 applies when the mean vectors form an angle on the interval (−180°,180°). The second case computes an angle on the interval (−180°,0°). The third case computes an angle on the interval (0°,180°). The fourth case gives 180° when the mean vectors point in opposite directions. The fifth case is when at least one of the mean vectors does not have a direction, which occurs when a mean vector has zero magnitude.

Equations 11.15 and 11.16 compute the Quantity component and the Allocation component of Mean Absolute Deviation for direction using the same logic that Eqs. 11.10 and 11.11 used for magnitude. Equation 11.17 computes the Direction MAD as the sum of its two components: Quantity and Allocation.

Figure 11.1 shows the results for the example data. The mean vector for **X** has magnitude 13 and angle 22.6°. The mean vector for **Y** has magnitude 10 and angle −36.9°. Thus the mean deviation has magnitude −3 and direction −59.5°. The magnitude's MAD is 2 plus the square root of 2, which is approximately 3.4. The direction's MAD is 90.

The computation of the mean for vector variables is different from how Chap. 8 computed the mean for interval variables. The equations in Chap. 8 for Correlation and Slope depend on the means for interval variables; therefore, those equations for Correlation and Slope from previous chapters do not apply to this chapter. Furthermore, Correlation and Slope indicate whether greater deviations are associated with larger **X** values but their logic does not apply to vector variables. The equations for Correlation and Slope in Chap. 8 fail to account for the fact that directions near 0° are closer to directions near 360° than to directions near 180°.

This chapter uses vectors that indicate the magnitude and direction of movement of a phenomenon, therefore a possible application is to compare datasets concerning wind. For example, if wind were blowing from southwest to northeast, then the vector would have positive components for *Xn* and *Xe*. However, some literature expresses wind in terms of a vector that indicates the source of the wind with a horizontal component *u* and a vertical component *v*, in which case wind that blows from southwest to northeast would have a negative components for *u* and *v* (Peng et al. 2013).

This chapter presents methods that make sense in a two-dimensional plane. If the plane is a map, then the geographic projection of the map must preserve direction so the methods of this chapter make sense. Users must think carefully when applying the methods of this chapter over regions so large that non-trivial distortion exists for some characteristics.

Two software packages exist to compute the concepts of this chapter. They are Vector Difference Analysis in Google Earth Engine (Tao and Bajracharya 2020) and VectorDeviation in R (Collins 2020).

11.2 Discussion Questions

1. What are some phenomena that vector variables describe?
2. If **Y** in Fig. 11.1 had a vector of magnitude 8 pointing south in the southwest pixel and a zero vector in the southeast pixel, then what would be the components of mean absolute deviation for magnitude and direction?
3. Under what circumstances does Eq. 11.14 produce a negative number of degrees?

References

Collins, E. (2020). *VectorDeviation*. https://github.com/skievanc/VectorDeviation

Peng, G., Zhang, H.-M., Frank, H. P., Bidlot, J.-R., Higaki, M., Stevens, S., & Hankins, W. R. (2013). Evaluation of various surface wind products with OceanSITES buoy measurements. *Weather and Forecasting, 28*, 1281–1303. https://doi.org/10.1175/WAF-D-12-00086.1.

Pontius Jr, R. G., Thontteh, O., & Chen, H. (2008). Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics, 15*, 111–142. https://doi.org/10.1007/s10651-007-0043-y.

Tao, S., & Bajracharya, A. (2020). *Wind Vector Comparison*. https://taoshiqi.users.earthengine.app/view/wind-vector-comparison

Chapter 12
Commandments to Avoid Deadly Sins

Abstract This chapter advises how to conduct science rigorously and to communicate effectively. The advice will help you to avoid pitfalls that are common in the profession. Hopefully, the advice will inspire you to produce innovative research that will lead to clear publications and scientific breakthroughs.

Keywords AUC · Kappa · Random · Significant

12.1 Text

I have developed these commandments by decades of teaching, advising, researching, reading, and writing. Furthermore, I have reviewed hundreds of manuscripts that others have submitted to journals. I see the same blunders so frequently that I keep a record of my recommendations for revisions so I can copy the recommendations into future reviews. I write this chapter so you can avoid these frequent pitfalls. If you follow this chapter's commandments, then your analysis and presentation will be clearer and more helpful than if you follow the profession's bad habits.

Commandment 1 is to design the research objective so that the results will be interesting and important, while you have no personal investment in any particular outcome. You must design your objective so if the results turn out one way, then the results are important; and if the results turn out a different way, then the results are also important. If your objective does not have this characteristic, then you should revise your objective. It is easy to convert your research objective from one that sets you up for sin into one that inspires insight. For example, if your initial research objective is to prove that one diagnostic method is more accurate than another method, then you should modify the objective to compare the behavior of one method vis-à-vis another method. If you are a rigorous scientist and have an interesting objective, then you and your audience will be interested in the results regardless of what the results show. If you desire to see particular results, then you will be tempted, perhaps unconsciously, to portray the results so they conform to your desires, which violates the principles of science. If you follow this commandment, then you do not need to worry about the results, which will relieve anxiety. Students and scientists are naturally interested in advancing their careers. Students want to

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1_12

graduate and graduates want to publish. There is no shame in that. We must advance in our careers to continue to do science. If you set up a situation where the results dictate whether you will advance, then you might never advance, or you will be tempted to bias your results, which will eventually hurt your career. Moreover, if you make your results conform to your preconceived notions or existing conventions, then you will never make breakthroughs. If you design your research objective according to this commandment, then you will increase your opportunity to discover something novel and publish innovative research. Emotional investment in the results causes unnecessary psychological stress. You should invest in methods that address important questions that will assure the results are interesting, regardless of what the results are.

Commandment 2 is to use metrics that relate to the research question. This advice is so obvious that you might wonder why any scientist would not follow this commandment. However, the commandment can be challenging in practice. For example, your initial research question might be "Which of various series for **Y** agrees the most with **X**?" That question might seem initially clear until the scientist realizes that the meaning of "agrees" is vague for an interval variable. Chapter 10 lists seven indices of agreement for interval variables. You must decide what metrics, if any, are relevant for your particular application. This might require that you change the research question to something clearer, such as "Which of various series for **Y** differs the least from **X**?" Even this revised question requires further refinement to define "differs" because various metrics measure various aspects of difference, as Chap. 8 illustrates for an interval variables. Scientists must avoid the sin of selecting a metric merely because the metric is popular in the profession. I have seen many cases where authors use a metric because of social tradition or because the available software computes the metric, while the metric does not relate to the research question. For example, Nash-Sutcliffe's E in Chap. 10 is popular in hydrology, but E uses one of many possible definitions of agreement. The particular definition of agreement that E uses might be irrelevant for a particular application. Some authors continue to report E due to mindless habit, despite well-documented conceptual problems with E (Criss and Winston 2008; Jain and Sudheer 2008). Another example is the kappa index of agreement, which is a problematic index of the agreement between two variables that show the same group of categories. Experts have analyzed the mathematical properties of kappa to recommend against the use of kappa (Foody 2020; Pontius and Millones 2011; Stehman and Foody 2019). However, other authors are slow to adopt the recommendation. I have asked many authors why they use E or kappa. I have never received an answer that the metric answers a particular research question. In most cases, the authors' responses indicate that the authors assume that readers expect E or kappa, which is a social reason that is unrelated to any particular research question. These social expectations derive from a dysfunctional tradition that has infected the profession like a contagious virus as one generation of scientists passes the tradition to the next. A similar situation exists for Root Mean Square Deviation (RMSD), which Chap. 10 describes. I usually ask authors why they use RMSD and they frequently tell me that RMSD measures the average distance between the $Y = X$ line and the (X,Y) points in a scatter plot.

However, Mean Absolute Deviation is the average distance while RMSD is not (Willmott et al. 2009; Willmott and Matsuura 2005). In this case, the authors knew their research question, but selected a metric that did not answer the question, usually because the authors have seen RMSD in other literature. Sophisticated readers appreciate literature that reports only the metric that answers the particular research question.

Commandment 3 is to decide whether a baseline is relevant; and if so, to use an appropriate baseline. For example, if a scientist tests a new model, then the baseline would be a default model. The default model could derive from an easily understandable naïve assumption. This default model is what I call the one-minute model, which is the model you could generate by thinking about the problem for 1 min. For example, if the application is to make a map that ranks pixels to predict urban growth, then a naïve assumption could be that urban growth expands from the edge of existing urban regions. I have seen cases where this naïve assumption produces a prediction that has a larger area under the TOC curve (AUC) than the AUC from a complicated machine-learning algorithm (Shafizadeh-Moghadam et al. 2021). Each of the indices of agreement in Chap. 10 uses a particular definition of baseline. The index is relevant only if the particular baseline is relevant. Some indices use a form of randomness as a baseline. Randomness is an appropriate baseline when the one-minute model assumes the process that generates the pattern is random. But many processes are obviously not random, while one minute's worth of thought could generate a more realistic non-random model. For example, urbanization does not happen randomly in space, whereas one minute's worth of thought could generate a model that assumes urbanization grows in proximity to previous urban places. This one-minute proximity model is likely to generate a baseline that has greater diagnostic ability than a model that assumes a random pattern. Some research objectives do not have a relevant baseline, such as when the purpose is to quantify temporal change, in which case the scientist should not use a baseline.

Commandment 4 is if you must decide on the acceptability of results, then define acceptability with respect to a particular research question for your case study. For example, scientists frequently want to know whether the data quality is acceptable to answer a particular research question, such as whether maps in a time series are sufficiently accurate to interpret the temporal difference as true change on the ground. Another example is whether validation of a method of prediction is sufficient to trust the method to predict the future. A sin is to use a universal rule to anoint results as acceptable, reasonable, good, excellent, or other subjective words. If you are going to use those words, then you must define those words for your particular application, which can be complicated. It is easier and clearer to refrain from those words. Some authors have attempted to define such words for various metrics such as the Area Under Curve, Percent Correct, Kappa, and other Indices of Agreement. Such universal rules cause more confusion than they resolve because they do not relate to any particular research question or specific case study. If a rule is universal, then the rule fails to consider the particular application by definition of the word universal. One common example is the unfortunate tradition in remote sensing where some misguided scientists define acceptable as any percent correct that is greater than 85%. This makes no sense because that criterion for acceptability is unrelated to any particular research question or case study. For example, 85% correct implies a 15% error. If the goal is to analyze change between two time points, then one must consider the error of the difference between the time points, not the accuracy of the data at the two individual time points. I have seen too many cases where the scientist claims that the data quality at each time point is acceptable because the percent correct is greater than 85%, and then the scientist interprets the difference from the start time to the end time as change on the ground. In many cases, the difference between the two time points is less than 15%, in which case, the error at the individual time points is larger than the difference between the time points. This makes me suspect that error could explain the temporal difference, but some authors miss this crucial point because the dysfunctional tradition in the profession has indoctrinated authors to think that 85% correct implies that the data are useful for any application.

Commandment 5 is to use inferential statistics if and only if you are sampling units that are meaningful to the research question. Inferential statistics examine how variation due to random sampling can influence the results. Inferential statistics play a dominant role in courses concerning introductory statistics, thus students leave the courses with the temptation to apply their limited tools to any situation, regardless of whether sampling plays a role. If your research does not use sampling, then you should not use inferential statistics. For example, if you have complete coverage for the region that constitutes your population, then you have a census of your region. If you have data for the population, then sampling uses less information than you have readily available, in which case there might be no reason for sampling. If you are not sampling, then you should not use inferential statistics, even when the software produces metrics that derive from inferential statistics. The p-value is a popular metric that inferential statistics generate. If the p-value is less than the alpha-level for a hypothesis test, then the observed statistic is significantly different than the hypothesized vaule of the parameter. I see frequently that authors inappropriately report p-values when the authors analyze data for the entire population. Many software packages compute p-values by default and some authors misinterpret a small p-value as meaning important or as indicating a large effect size. Larger numbers of observations will cause smaller p-values. Populations tend to have a large number of observations, which causes small p-values that lead a scientist to anoint the results as significant, even when the effect size is so small as to have no practical importance. On a related note, use the word “significant” if and only if you are referring to a p-value computed with inferential statistics. Realize that “significant” in inferential statistics does not necessarily mean important. Perhaps the most common application of an arbitrary universal rule is the tradition of using a threshold of alpha equal to 0.05 to denote results as significant or not. There is nothing universally magical about the traditional threshold of 0.05, but some scientists focus on whether their p-value is less than 0.05 in a desire to call the results significant. This book does not use inferential statistics because other literature, including textbooks for introductory statistics classes, describes concepts and equations for inferential statistics (Olofsson et al. 2014). If sampling applies to your analysis, then you must account for the sampling design when giving summary statistics. Stratified random sampling can be helpful to obtain information efficiently. An intelligently designed stratified random sample frequently has the number of samples in each stratum not proportional to the size of the stratum. It makes sense to concentrate the samples in the strata where you invest the least cost to obtain the most helpful information. If the number of samples in each stratum is not proportional to the size of the stratum, then scientists must use appropriate equations, such as those in Chap. 5, to convert the sample data to unbiased estimates of the population.

Commandment 6 is to present the data and results visually in various forms. A visual assessment allows the human brain to identify relationships that any particular metric might miss. You must consider several ways to visualize the data because any single visual form will highlight some aspects and hide other aspects. The maps at the top of Fig. 7.1 illustrate the concept when analyzing land change. Persistence typically accounts for most of the spatial extent in a time series of land categories, in which case the maps at the various time points fail to reveal change clearly. Thus, Fig. 7.1 shows persistence as gray so the maps highlight the categories that lose and gain during each time interval. This format has been enlightening for many applications (Aldwaik and Pontius Jr 2012, 2013; Enaruvbe and Pontius Jr 2015; Pontius Jr 2019; Pontius Jr et al. 2011, 2013, 2017, 2018; Varga et al. 2019; Xie et al. 2020). You must design the graphic for the intended purpose and not allow the default settings in the software to dictate an inappropriate format. For example, the scatter plots in Chaps. 8 and 9 have identically formatted axes and include the $Y = X$ line because the purpose is to see how Y differs from X. The default settings in most software packages do not have those characteristics, thus the user must customize the figures. Chapter 9 concerning sea surface temperature illustrates the data in the form of maps, scatter plots, and components of difference, because each form highlights aspects that another form hides. The human brain can see several aspects of patterns that tables of numbers fail to convey. Visual examination sometimes identifies patterns that derive from data problems, which are essential to know before digging into detailed analysis. My professor advised me decades ago that one of the first steps in analysis is to plot the data and look at it. This advice has proved remarkably effective.

Commandment 7 is to follow the Three Ones Principle, which is to use exactly one word to mean exactly one concept in exactly one piece of literature. Readers are baffled when one piece of literature uses more than one word to mean one concept. A typical example is the word “significant”. I have seen literature that uses the word significant in some places to refer to a small p-value and in other places to mean important. But, a small p-value is not equivalent to important. Another problematic word in the literature is “random”. Use random if and only if you mean mathematically random. Mathematical randomness is a very specific pattern, which usually requires a random number generator for practical implementation. I hear people use the word random in casual conversation. Sometimes those people seem to use random seems to refer to a haphazard procedure, meaning the procedure lacks rigor. Other times, people use random to refer to a pattern that they do not understand. Yet other times random means variation for which a model does not account. Frequently the meaning of the word random is unclear, perhaps even to the people who use the word. However, random has an extremely specific meaning in mathematics. A random sample selects members such that each member of a population has an identical probability of selection.

Commandment 8 is to write sentences that have a simple grammatical structure in the active voice. The best advice that my doctoral advisor, Dr. Charles Hall, gave me is: Put the subject at the beginning of the sentence, then put the verb immediately following the subject, then finish the sentence. Professor Hall says the advice will clarify 85% of language ambiguities. My experience confirms his claim. Scientific writing should differ from writing in poetry or novels. I become confused when an author uses complex elaborate sentences, and I read in my native language, English. Much of your audience might not be using their native language to read your paper, so keep the grammatical structure simple and the language straight forward. As an added benefit, your thoughts become clearer to you when you force yourself to write sentences that have a straightforward structure. Consider the sentence "Contradicting the hypothesis, it was found that the mean of **Y** was greater than the mean of **X**." This sentence does not begin with the subject and verb. The subject is a vague pronoun. The verb is the past tense, while the present tense is more appropriate. The passive voice fails to communicate who did what, meaning whether the authors found something or whether other scientists found something. A clearer sentence would begin "Our results show that the mean of **Y** is greater than the mean of **X**, which contradicts our hypothesis."

Commandment 9 is to be kind and courageous. You must appreciate the emotional, psychological, and social aspects of science. Scientists have feelings just like any other humans, so you must treat your colleagues with kindness to earn their trust and to enhance collaboration. If you think another scientist has made blunders, then you should ask why the scientist did what the scientist did and offer a more enlightening alternative. If you criticize others harshly without offering any alternative, then you will alienate your colleagues, which will hinder your efforts to disseminate better ideas. You must also treat yourself with kindness when you realize that you have made mistakes. If you have a curious mind, then you will realize during your career that some of your earlier thoughts were flawed. This is a natural part of the learning process. If you have made previous blunders, as I have, then you should admit them openly and describe how your thought processes have evolved. If you think none of your previous work was flawed, then I suspect your learning has stopped. You must also have the courage to blaze a path that makes sense to you, even when you encounter resistance from others. You must have the courage to publish literature that you are convinced is correct, even when your literature does not follow popular conventions. Popular conventions might prevent you from both discovery and innovation. You must resist the temptation to report metrics simply because you have seen others report the same metrics. If those metrics are flawed, then you will repeat the mistakes of others and encourage the next generation to continue the mistakes. Report metrics if and only if the metrics make sense to you. You should test your thoughts by explaining the metrics to your colleagues. When I have had difficulty in explaining a metric to others, I have later realized that the metric was conceptually flawed. You must work with colleagues to develop clear ideas. To obtain the cooperation of others, you must treat others with kindness. After you earn their trust, they will be more likely to listen when you challenge conventional ideas. I hope this book inspires you to use metrics that shed light on your scientific path. If you use the metrics in this book, then you and your audience are likely to gain deep insight. I hope my book has treated you with kindness and will inspire you to have the courage to use *Metrics* that Make a Difference.

12.2 Discussion Questions

1. What would motivate a scientist to commit one of the deadly sins of this chapter?
2. Some authors recommend universal rules to anoint particular values of a metric as excellent, good, acceptable, or poor. Do you endorse such rules? Why or why not?
3. How do the meanings of the word "significant" and "random" in casual conversation differ from their meanings in scientific communication?
4. How did you or other audience members respond when you saw a scientist commit one of the deadly sins of this chapter? What was the reason for the response or lack of response? How will you respond next time you see a colleague commit a sin?

References

Aldwaik, S. Z., & Pontius Jr, R. G. (2012). Intensity analysis to unify measurements of size and stationarity of land changes by interval, category, and transition. *Landscape and Urban Planning, 106*, 103–114. https://doi.org/10.1016/j.landurbplan.2012.02.010.

Aldwaik, S. Z., & Pontius Jr, R. G. (2013). Map errors that could account for deviations from a uniform intensity of land change. *International Journal of Geographical Information Science, 27*, 1717–1739. https://doi.org/10.1080/13658816.2013.787618.

Criss, R. E., & Winston, W. E. (2008). Do Nash values have value? Discussion and alternate proposals. *Hydrological Processes, 22*, 2723–2725. https://doi.org/10.1002/hyp.7072.

Enaruvbe, G. O., & Pontius Jr, R. G. (2015). Influence of classification errors on intensity analysis of land changes in southern Nigeria. *International Journal of Remote Sensing, 36*, 244–261. https://doi.org/10.1080/01431161.2014.994721.

Foody, G. M. (2020). Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. *Remote Sensing of Environment, 11*. https://doi.org/10.1016/j.rse.2019.111630.

Jain, S. K., & Sudheer, K. P. (2008). Fitting of hydrologic models: A close look at the Nash–Sutcliffe index. *Journal of Hydrologic Engineering, 13*, 981–986. https://doi.org/10.1061/(ASCE)1084-0699(2008)13:10(981).

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment, 148*, 42–57. https://doi.org/10.1016/j.rse.2014.02.015.

Pontius Jr, R. G. (2019). Component intensities to relate difference by category with difference overall. *International Journal of Applied Earth Observation and Geoinformation 77*, 94–99. https://doi.org/10.1016/j.jag.2018.07.024

Pontius Jr, R. G., Castella, J.-C., de Nijs, T., Duan, Z., Fotsing, E., Goldstein, N., Kok, K., Koomen, E., Lippitt, C.D., McConnell, W., Mohd Sood, A., Pijanowski, B., Verburg, P., & Veldkamp, A.T. (2018). Lessons and Challenges in Land Change Modeling Derived from Synthesis of Cross-Case Comparisons, in: Behnisch, M., Meinel, G. (Eds.), Trends in Spatial Analysis and Modelling, Geotechnologies and the Environment. Springer International Publishing, Cham, pp. 143–164. https://doi.org/10.1007/978-3-319-52522-8_8

Pontius Jr, R. G., Gao, Y., Giner, N., Kohyama, T., Osaki, M., & Hirose, K. (2013). Design and Interpretation of Intensity Analysis Illustrated by Land Change in Central Kalimantan, Indonesia. Land 2, 351–369. https://doi.org/10.3390/land2030351

Pontius Jr, R. G., Krithivasan, R., Sauls, L., Yan, Y., & Zhang, Y. (2017). Methods to summarize change among land categories across time intervals. *Journal of Land Use Science 12*, 218–230. https://doi.org/10.1080/1747423X.2017.1338768

Pontius Jr, R. G., & Millones, M. (2011). Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing, 32*, 4407–4429. https://doi.org/10.1080/01431161.2011.552923.

Pontius Jr, R. G., Peethambaram, S., & Castella, J.-C. (2011). Comparison of Three Maps at Multiple Resolutions: A Case Study of Land Change Simulation in Cho Don District, Vietnam. *Annals of the Association of American Geographers 101*, 45–62.

Shafizadeh-Moghadam, H., Minaei, M., Pontius Jr, R. G., Asghari, A., & Dadashpoor, H. (2021). Integrating a forward feature selection algorithm, random forest, and cellular automata to extrapolate urban growth in the Tehran-Karaj Region of Iran. *Computers, Environment and Urban Systems, 87*, 101595.

Stehman, S. V., & Foody, G. M. (2019). Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment, 231*, 111199. https://doi.org/10.1016/j.rse.2019.05.018.

Varga, O. G., Pontius Jr, R. G., Singh, S. K., & Szabó, S. (2019). Intensity analysis and the figure of Merit's components for assessment of a cellular automata – Markov simulation model. *Ecological Indicators, 101*, 933–942. https://doi.org/10.1016/j.ecolind.2019.01.057.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*, 79–82. https://doi.org/10.3354/cr030079.

Willmott, C. J., Matsuura, K., & Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment, 43*, 749–752. https://doi.org/10.1016/j.atmosenv.2008.10.005.

Xie, Z., Pontius Jr, R. G., Huang, J., & Nitivattananon, V. (2020). Enhanced intensity analysis to quantify categorical change and to identify suspicious land transitions: A case study of Nanchang, China. *Remote Sensing, 12*, 3323. https://doi.org/10.3390/rs12203323.

Glossary

Abundance size of sum of Hits and Misses

active situation when a category's loss intensity or gain intensity is greater than the uniform intensity at the category level for Intensity Analysis

Allocation component of difference that is either the sum of Exchange and Shift components for a categorical variable or two times the minimum of the sum of negative deviations and the sum of positive deviations for an interval variable

AUC Area Under the Curve of the Relative Operating Characteristic, which is equal to the ratio where the numerator is the area under TOC curve in TOC's parallelogram and the denominator is the area of TOC's parallelogram

avoid situation when the transition intensity from a particular losing category to a gaining category is less than the gaining category's uniform transition intensity at the transition level for Intensity Analysis

bias mean **Y** minus mean **X** when **X** and **Y** show the same interval phenomenon

binary variable variable that shows exactly two distinct states, such as Presence or Absence

Boolean variable variable that shows exactly two distinct states, such as Presence or Absence

calibration procedure to use data to determine the parameters in a model

categorical variable variable that shows distinct states

Cellular Automata algorithm that influences how a cell changes based on the cell's neighbors

component portion of difference that derives from Quantity, Exchange, Shift, Allocation, across strata or within strata

Composite Matrix square contingency table for observations that belong to more than one category

Correct Rejection observation for which diagnosis and truth are Absence for a category

Correlation index ranging from −1 to 1 that measures strength and sign of linear association between two interval variables

deviation difference for an interval variable

R. G. Pontius Jr, *Metrics That Make a Difference*, Advances in Geographic Information Science, https://doi.org/10.1007/978-3-030-70765-1

dormant situation when a category's loss intensity or gain intensity is less than the uniform intensity at the category level for Intensity Analysis

Exchange component of difference that derives from when at least one observation transitions from category *i* to category *j* while simultaneously at least one other observation transitions from category *j* to category *i*

extent collection of observations of interest, meaning the population

False Alarm observation for which diagnosis is Presence and truth is Absence, also known as False Positive and Type I error

Hit observation for which diagnosis and truth are Presence, also known as a True Positive

Inferential Statistics quantitative analysis that uses a sample from the population to make inferences or test hypotheses concerning a population parameter

intensity ratio where the numerator is the size of difference and the denominator is the size where the difference could have possibly occurred

Intensity Analysis analytical framework that describes three levels of difference between two variables that show the same categories

interval variable variable for which addition and subtraction makes sense

Kappa convoluted index of agreement between two categorical variables that show the same set of categories

MAD Mean Absolute Deviation, which is the mean absolute vertical distance between the $Y = X$ line and the points in a scatter plot where **X** and **Y** are interval variables that show the same phenomenon

Miss observation for which diagnosis is Absence and truth is Presence for a category, also known as a False Negative and Type II error

net size of gain minus size of loss, thus net can be negative

observation record in a database

pattern characteristic of data

Prevalence number of Presence observations in reference data divided by number of observations in extent

process phenomenon that generates patterns in data

population collection of all possible observations of interest

Quantity component of difference that measures absolute net difference between **X** and **Y**

rank variable variable that uses whole numbers to rank observations from least to greatest or greatest to least

Recall Hits divided the sum of Hits and Misses, also known as Sensitivity

resolution most detailed characteristic of the observations, such as the size of the smallest spatial unit of observation or the shortest time interval between observations

RMSD Root Mean Squared Deviation, which is the square root of the mean squared deviation between observations of interval variables **X** and **Y** where both variables show the same phenomenon

ROC Relative Operating Characteristic, also known as Receiver Operating Characteristic, which shows a relationship between a binary variable and a rank variable; ROC is less informative than the Total Operating Characteristic (TOC)

R-squared proportion of the variance in **Y** for which the variance in **X** accounts based on a particular mathematical function for **Y** as a function of **X**; if the function is linear, then R-squared is equal to the square of Pearson's correlation coefficient.

sample subset of the population for which data exist

scale word that can indicate various concepts such as spatial resolution, spatial extent, temporal resolution, temporal extent, or categorical detail

Sensitivity Hits divided the sum of Misses and Hits, also known as Recall

Shift component of difference that derives from when at least one observation transitions from category *i* to category *j* while simultaneously at least one other observation transitions from category *j* to category *k* where *j* differs from *k*

significant situation when the p-value is less than the alpha-level for a hypothesis test using inferential statistics. Statistical significance does not necessarily imply practical importance.

size amount in a quantitative analysis, such as the area of a category

Specificity Correct Rejections divided the sum of False Alarms and Correct Rejections

strata plural of stratum

stratum subset used to partition the population

target situation when the transition intensity from a particular losing category to a gaining category is greater than the gaining category's uniform transition intensity at the transition level for Intensity Analysis

TOC Total Operating Characteristic, which shows a relationship between a binary variable and a rank variable

transition change through time from one category to a different category

validation procedure to measure how a model's output relates to the corresponding reference data that were not used for calibration