Ученые записки Таврического национального университета имени В.И.Вернадского
Серия «География». Том 26 (65). 2013 г. № 1, С. 12-22.

*УДК 519.257*

# КАРТОГРАФИЧЕСКОЕ ПОНИМАНИЕ СТАТИСТИЧЕСКИХ ДАННЫХ

***Бакли Эйлин***

***Esri, Рэдлэндс, США***
***E-mail:***abuckley@esri.com
***Перевод: Дядюн В.Ю., ЧАО "ЕСОММ"***

Многие карты изображают статистические или числовые данные. Понимание статистических данных и методов их картографирования является ключом к построению интуитивно понятной статистической карты. Статья исследует вопросы картографирования статистических данных.
***Ключевые слова:*** качественные, количественные, пространственно-интенсивные, пространственно-экстенсивные, метод хороплет, пространственная статистика

## 1. КАЧЕСТВЕННЫЕ И КОЛИЧЕСТВЕННЫЕ

По сути, карты отображают только два вида данных: качественный и количественный. Качественные данные изменяются между различными типами вещей. Количественные данные передают величину. Хотя любой вид данных может быть передан на карте точками, линиями, полигонами и растровыми ячейками, методы, используемые для нанесения на карту этих двух видов данных, несколько отличаются.

Категориальные различия в качественных данных могут быть показаны символами, различающимися по цветовому оттенку (например, красный, зеленый, синий) и форме (например, окружности, квадраты, треугольники). Количественные данные также могут быть эффективно отображены через изменения параметров символики таких, как ориентированность и структура пространственного образца, однако оттенки цвета, форма, яркость и размер символа используются чаще всего, поскольку наиболее просты для понимания..

Был разработан ряд картографических методов, объединяющих географические объекты и символы. Метод хороплет использует яркость для отображения полигонов. Карты пропорционального символа отображают результаты в виде точек, которые различаются по размеру, основываясь на связанных с ними значениях.

Поскольку большая часть статистических данных является количественной по природе, эта статья фокусируется на картографировании количественных данных. Однако, чтобы надлежащим образом картографировать количественные данные, их нужно понять. Не все методы одинаково применимы.

Демографические данные тому пример. Они показывают статистические характеристики численности народонаселения и являются одним из наиболее общих видов данных, показываемых на статистических картах. Демографические данные, которые могут включать данные о расе, поле, возрасте, статусе занятости и иные показатели, сведены в таблицу по счетным блокам таким как области, зоны

переписи населения, области ZIP кодов или школьные округи. Таблицы включают сумму числа объектов, например лиц, домашних хозяйств, жилых домов или студентов внутри счетных блоков. Они также могут включать характеристики этих объектов, например возраст, расу и доходы, чтобы описать людей или возраст и тип жилого строения.

Сумма числа объектов и характеристики объектов могут быть использованы для получения измерений, выражающих реферирование (например, среднее, медиана) или взаимозависимости (например, плотности, пропорции). Таблицы и производные значения для счетных блоков предположительно одинаковы в любой точке области и меняются на границах блока (например, они не перетекают из одного блока в другой).

Ландшафтные показатели для водоразделов и показатели налогообложения для кадастровых участков являются примерами данных, собранных для блока в целом и о которых можно предположить, что они равномерно распределены по счетной единице и изменяются на ее границах. Прежде, чем наносить такие данные на карту, кроме определения того, обладают ли оцениваемые данные такими характеристиками, нужно выяснить еще следующее.

## 2. ПРОСТРАНСТВЕННО-ЭКСТЕНСИВНЫЕ И ПРОСТРАНСТВЕННО-ИНТЕНСИВНЫЕ ДАННЫЕ

Нужно также учитывать зависимость подлежащих нанесению на карту статистических данных от размера счетного блока. Суммы числа объектов и величины типа “всего” вместе с измерениями, например площадью и периметром, являются сводными статистиками для счетного блока и верны лишь когда представляют блок в целом. Говорят, что такие статистики пространственно-экстенсивные. Статистика является суммой свойств элементов, составляющих счетный блок. Например, суммарные величины типа “всего” являются суммой пунктов, просуммированных в счетном блоке. Периметр является суммой длин линейных сегментов, образующих границу счетной единицы. Если размер счетной единицы меняется, статистики также меняются.

Для сравнения, значения, такие, как плотность населения или уровень заболеваемости раком могут описать любую часть счетной единицы (если счетная единица предполагается однородной). Такие статистики являются пространственно-интенсивными и не зависит от размера счетного блока. Если разделить счетную единицу, то статистика не изменится.

Пространственно-интенсивные данные могут быть получены на основе пространственно-экстенсивных. Например, деление сумм объектов на площадь формирует плотность или деление суммы для одной счетной единицы на сумму сумм объектов для всех счетных единиц формирует пропорцию.

Чтобы лучше это понять, взгляните на рис. 1. Данные для пяти счетных блоков показывают число людей, площадь и плотность населения для каждой единицы. Повторное вычисление значений, основанное на произвольном разбиении исходных счетных блоков, выявило, что пространственно-интенсивные измерения, например

плотность, не зависят от размера области, тогда как пространственно-экстенсивные переменные, например площадь или сумма объектов, являются пространственно-зависимыми.

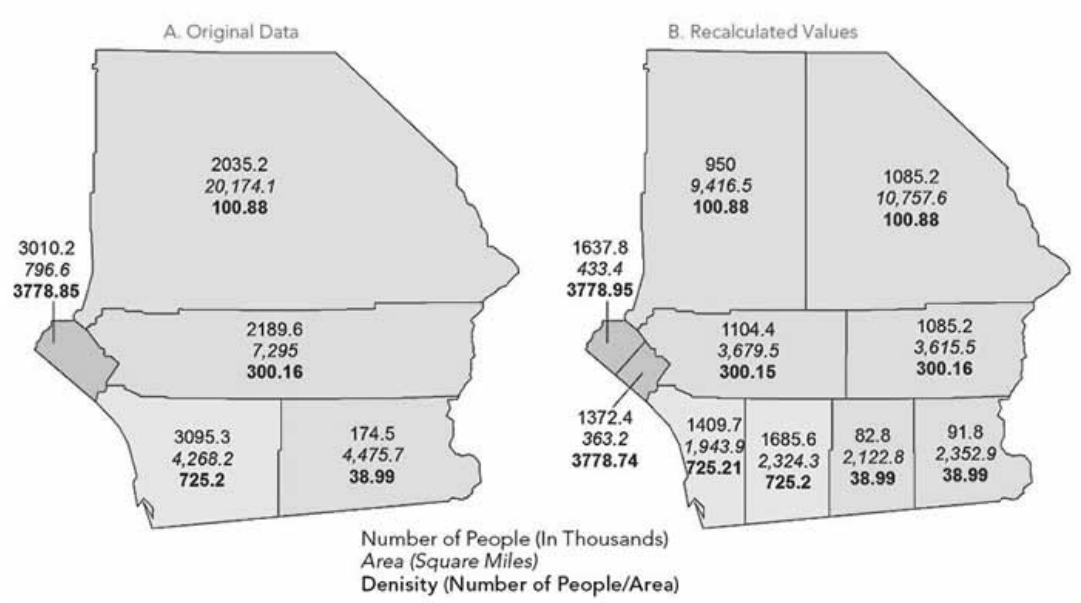


Рис. 1-А показывает статистики для числа лиц, площади и плотности (люди/площадь) для пяти счетных единиц. Рис. 1-В показывает блоки произвольно разделенные на 10 новых единиц.

Можно перевычислить все статистики предполагая, что исходные суммы объектов однородны в пределах счетной единицы, что является одним из предположений для демографических данных обсуждавшийся ранее. Площадь может быть легко перевычислена, как процент от старой площади которую занимает новая площадь (новая площадь/старая площадь). Для вычисления новой суммы старая сумма умножается на процент площади новой счетной единицы. Это новое значение будет правильным только в предположении, что число людей распределено случайно внутри счетного блока. Однако, повторное вычисление плотности дает тоже значение, так как сумма объектов прямо пропорционально зависит от площади.

Карта на рис. 2 показывает картографирование данных по методу хороплет. Когда суммы объектов символизируются посредством яркости (заметим, что более темное понимается всегда как содержащее больше значений), карта повторно вычисленных значений существенно отличается от карты с оригинальными значениями.

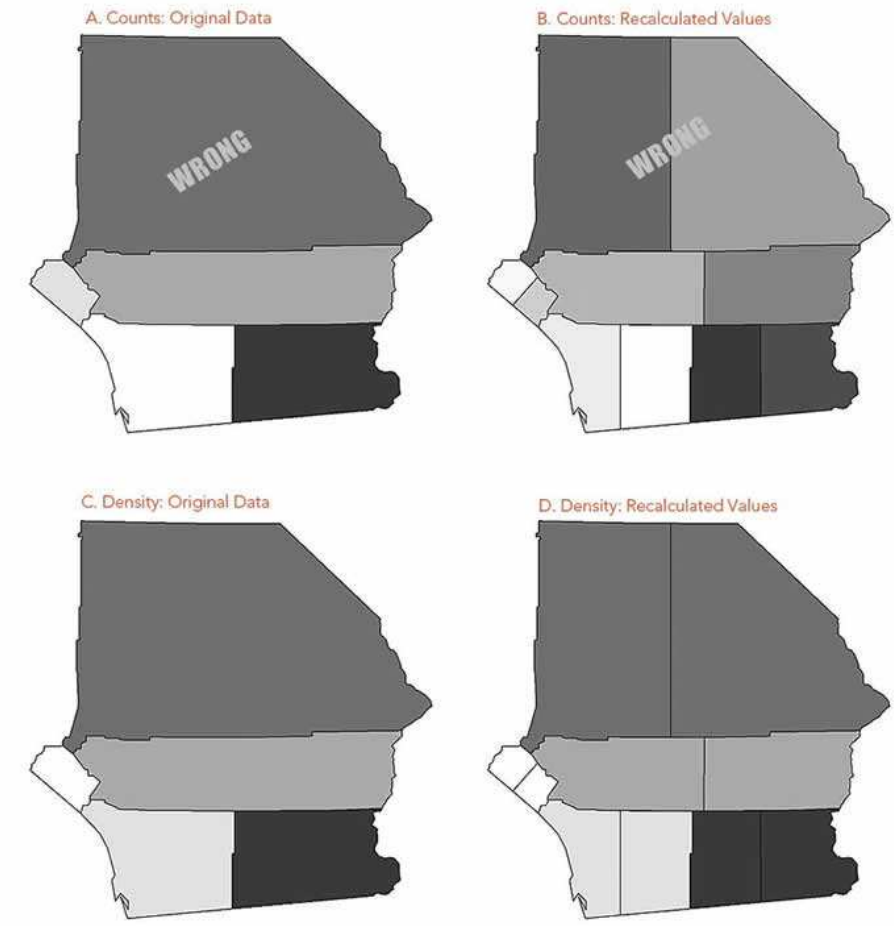


Рис. 2. Правильно будет использовать метод хороплет для картографирования плотности, но не сумм объектов.

Это нарушает предположение о том, что значения в счетных блоках распределены равномерно внутри них. Однако, при картографировании плотности распределения выглядят в точности такими же. Произвольное деление счетных единиц показывает свойства пространственной интенсивности и экстенсивности данных, но это, скорее всего, нежелательно.

Давайте взглянем на актуальные данные. Данные неравномерно распределены внутри счетной единицы, как и в случае с большинством площадных данных. Исходные счетные единицы (показанные на рис. 3) имеют отношение к областям и далее поделены на зоны переписи населения. Нанесение плотности населения этих зон на карту показывает, что в рамках всей области население сконцентрировано на юго-западе. Однако картографирование населения по областям скрывает эту вариацию в распределении.

Существует другая проблема с картографированием сумм объектов и суммарных величин, и суммарных величин типа “всего” и других пространственно-экстенсивных данных для областей по методу хороплет. Однородные распределения будут скрыты. Карты на рис. 4A, 4B и 4C показывают картографированные данные вначале как однородное распределение, затем как две карты-хороплеты отображающие суммы объектов и плотность. Сумма изменяется в широких пределах для области, что ведет к промежутку яркости на карте на рис. 4B. Хотя плотность одинакова для всех областей, эта вариация дает ложное понимание метода распределения объектов внутри областей. Для сравнения, карта на рис. 3C имеет ту же плотность для каждой счетной единицы. Отсутствие вариации яркости между счетными единицами дает верное восприятие распределения объектов.

Рис. со 2 по 4 показывают очень важный принцип: суммы и итоговые величины, а также другие пространственно-экстенсивные данные не должны отображаться методом хороплет.

Почему?

Потому, что этот метод не совсем точно передает природу данных. Отображение пространственно-экстенсивных данных с использованием метода хороплет скрывает концентрацию объектов в областях, потому что он предполагает равномерное распределение, как показано на картах на рис. 2. Метод хороплет также скрывает равномерные распределения, как показано на картах на рис. 4. Хотя это всего лишь один пример использования метода отображения, который не соответствует типу данных, однако он встречается достаточно часто.

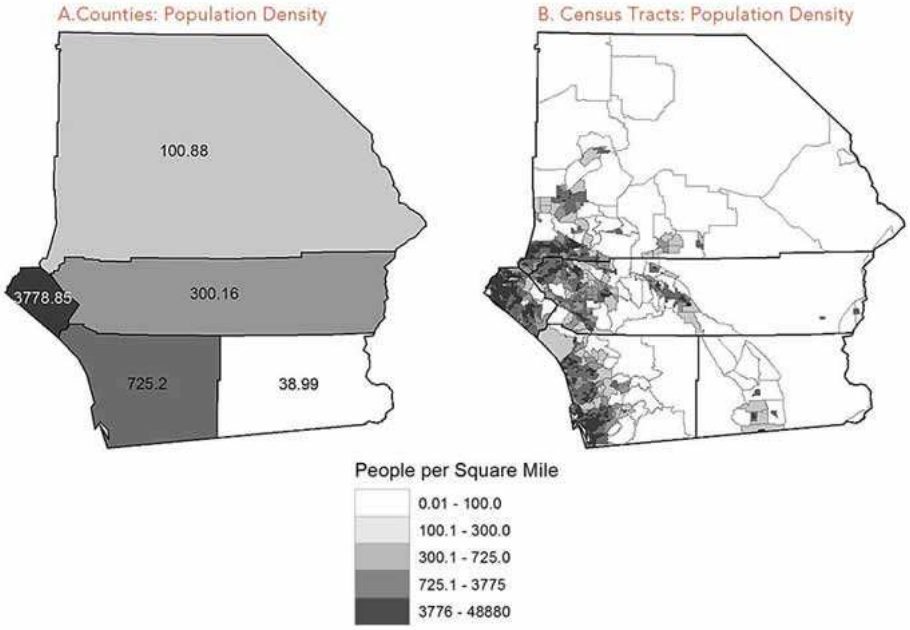


Рис. 3. Картографирование плотности населения для зон переписи населения (B) показывает, что люди сконцентрированы на юго-западе – факт, не видный из плотности населения по областям (A).

## 3. НОРМАЛИЗАЦИЯ ИЛИ СТАНДАРТИЗАЦИЯ ДАННЫХ

Теперь, когда понятно, что метод картографирования должен соответствовать природе данных, следующим шагом будет поиск правил работы с данными с таким расчетом, чтобы они соответствовали используемому методу.

Чтобы исправить проблемы возникшие при картографировании сумм объектов по методу хороплет, нужно сконвертировать данные к такому виду, чтобы его можно было показывать яркостью для областей. Это часто необходимо для данных, представленных в виде точек, линий или растров и с помощью других методов отображения, таких как пропорциональный круг.

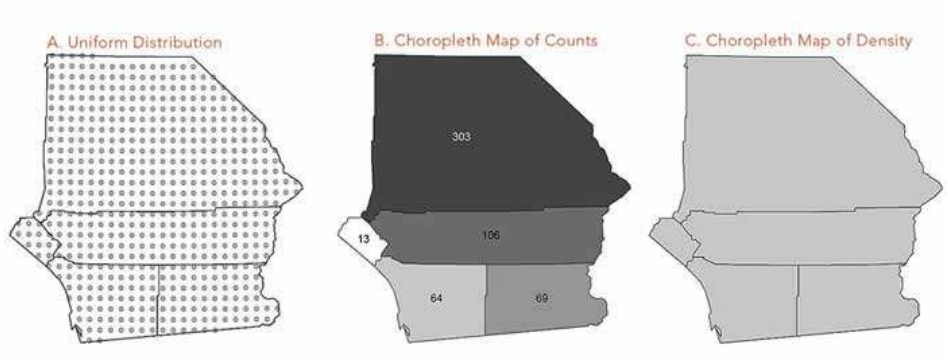


Рис. 4. Равномерное распределение (A) скрыто методом картографирования хороплет для показа пространственно-экстенсивных данных, таких как суммы объектов (B), а не статистик для пространственно-интенсивных данных, таких как плотность (C).

Это выполняется путем нормализации или конвертации данных. Эти два термина, часто используемые как синонимы, немного отличаются. Нормализация данных отображает все числовые значения на диапазон от нуля до единицы. Стандартизация преобразует данные таким образом, чтобы они имели нулевое среднее и единичную дисперсию. Оба метода имеют свои недостатки. Если набор данных имеет выбросы, нормализация будет отображать данные на очень малый интервал. При использовании стандартизации предполагаем, что данные были получены с определенным средним и стандартным отклонением, хотя, это может быть и не так.

## 4. МЕТОДЫ ПОЛУЧЕНИЯ ПОДХОДЯЩИХ ИЗМЕРЕНИЙ

Картографы часто используют термин «полученные данные», называя им данные, которые преобразованы путем нормализации и стандартизации и поэтому могут сравниваться. Преобразования часто используются в картографии, например, отношения, пропорции, проценты и плотность.

Важно различать пространственно-интенсивные и пространственно-экстенсивные измерения. Плотность является пространственно-экстенсивным измерением. Пропорция, порожденная делением числа пунктов в счетной единице

на общее количество пунктов, является пространственно-экстенсивной, так как число на блок было разделено на константу (общее количество объектов). Для производных значений, таких как пропорции, проценты и отношения, полученные числа могут быть только истинными для всей счетной единицы, а не его части. Для счетных единиц особенной важности (например, областей) картографирование пропорции значения отведенного на каждую единицу не должно использовать метод хороплет. В таких случаях, может быть, лучше использовать градуированный символ.

Рис. 5 показывает карты некоторых типов полученных данных. На рис. 5A приведены две статистики в исходных данных, которые были использованы при вычислениях – число студентов и учителей в каждой счетной единице. Площадь каждой счетной единицы может быть вычислена в ГИС. С использованием формул из таблицы 1 были созданы карты, которые показывают плотность учителей (5B), процент учителей (5C) и отношение студентов к учителям (5D) для каждой счетной единицы. Эти очень различные карты могут быть использованы для ответа на очень разные вопросы.

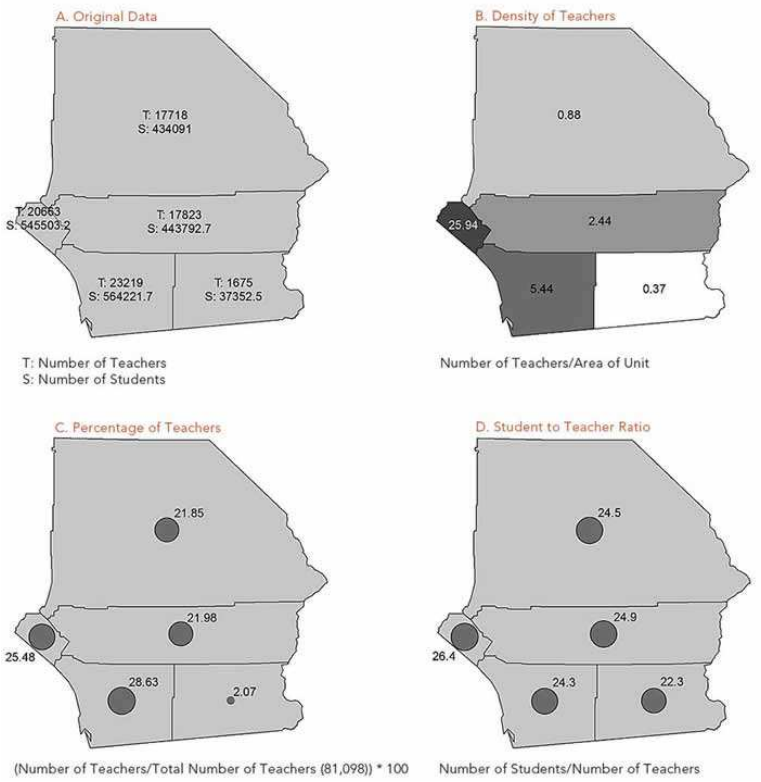


Рис. 5. Исходные данные включают суммы учителей и студентов. Площадь счетных единиц была вычислена в ГИС. Эти пространственно-экстенсивные измерения могут быть преобразованы в пространственно-интенсивные, которые могут быть отображены методом хороплет. Примеры включают плотности (B), проценты (C) и отношения (D). (Источник данных: California Ed-Data website).

Таблица 1

| Преобразование | Операция |
|---|---|
| Коэффициенты выражают отношение одного наблюдения к другому. | Ratio or rate = $n_a/n_b$ |
| Пропорции выражают отношения одного наблюдения ко всем наблюдениям. | Proportion = $n_a/N$ |
| Проценты выражают тоже самое что пропорции, но используя значения из промежутка от 1 до 100. | Percentage = $n_a/N * 100$ |
| Плотности выражают связь между наблюдением и размером счетной единицы. | Density = $n_a/A$ |

Часто используемые картографические преобразования, вычисленные с использованием следующих операций, при этом не является числом наблюдений в одной категории, но является числом наблюдений в другой категории, N является общим числом категорий и A является площадью счетной единицы.

Например, знание плотности учителей помогает ответить на вопросы как, где сконцентрированы учителя? Это может быть полезно, если нужно провести встречу в месте, которое позволит свести к минимуму перемещение для большинства учителей. Знание процента учителей в каждой счетной единице помогает ответить на такие вопросы, как: сколько учителей из всех находится в каждой счетной единице? Это было бы полезно для выделения средств на учителей для школьных принадлежностей.

Одной из проблем является то, что производные значения могут скрывать характер данных, используемых в расчетах. Например, карта на рисунке 3 может скрыть тот факт, что не все учителя работают полный рабочий день. Два учителя-совместителя могут рассматриваться как два учителя, но вместе являются эквивалентом одного учителя на полной ставке. Этот аспект данных не учитывается, если только отображаются учителя на полной ставке, а не общее число учителей.

Кроме того, величины, которые не являются сопоставимыми, не должны быть использованы для расчета коэффициентов. Например, не стоит вычислять (или картографировать) число учителей в школе, если все школы не являются примерно одинаковыми по размеру. Для того, чтобы коэффициент имел смысл, школы должны быть сопоставимыми.

## 5. РЕЗЮМЕ

Большее понимание природы статистических данных, используемых для целей картографирования поможет лучше понять методы, которые могут быть использованы для сопоставления карт. В конечном счете целью является согласование соответствующих данных и наиболее эффективных методов таким

образом, чтобы карта могла быть легко, быстро и правильно интерпретирована читателями.

## 6. ОБ АВТОРЕ

Эйлин Бакли является профессиональным картографом с более чем 25-летним опытом работы. Последние 10 лет она была в Esri, где сосредоточилась на разработке наилучшей практики картографирования в ArcGIS и совместного использования этих методов с пользователями. В дополнение ко многим интернет-ресурсам, она написала множество статей для Esri и других профессиональных изданий. Эйлин является одним из авторов учебника Использование карты: Чтение, анализ и интерпретация (Map Use: Reading, Analysis, and Interpretation). Она также представляет по всему миру широкий спектр тем, касающихся картографирования и ГИС. Эйлин имеет три степени по географии: бакалавра в университете Вальпараисо в штате Индиана, мастера в рамках совместной программы Университета Индианы и Университета штата Мичиган, и доктора Университета штата Орегон. Она также принимала участие в магистерской программе по ГИС в Университете Рэдландс с момента ее создания.

## References

1. Brewer, Cynthia A. 2006. "Basic Mapping Principles for Visualizing Cancer Data Using Geographic Information Systems (GIS)," www.ajpmonline.org/article/S0749-3797%2805%2900358-2/fulltext.
2. Cote, Paul. Effective Cartography: Mapping with Quantitative Data, www.gsd.harvard.edu/gis/manual/normalize/.
3. Kimerling, A. Jon, Aileen R. Buckley, Phillip C. Muehrcke, and Juliana O. Muehrcke. 2011. Map Use: Reading, Analysis, Interpretation, Seventh Edition. Redlands, CA: Esri Press, 581 pages.
4. Longley, Paul A., Michael F. Goodchild, David J. Maguire, and David W. Rhind. 2011. Geographic Information Systems and Science, Third Edition, New York: Wiley, Chapter 4.
5. Pitzl, Gerald. 2004. Encyclopedia of Human Geography. "Choropleth maps."
6. Robinson, Arthur H., Joel L. Morrison, Phillip C. Muehrcke, A. Jon Kimerling, and Stephen C. Guptill. 1995. Elements of Cartography, Fifth Edition. New York: John Wiley & Sons, Inc., 674 p.
7. Saitta, Sandra. "Standardization vs. normalization," Data Mining Research blog.

**Баклі Е. Картографічне розуміння статистичних даних / Е. Баклі** // Вчені записки Таврійського національного університету імені В. І. Вернадського. Серія: Географія. – 2013. – Т.26 (65). – № 1 – С. 12-22.

Багато карт зображують статистичні або числові дані. Розуміння статистичних даних і методів їх картографування є ключем до побудови інтуїтивно зрозумілої статистичної карти. Стаття досліджує питання картографування статистичних даних.

***Ключові слова***: якісні, кількісні, просторово-інтенсивні, просторово-екстенсивні, метод хороплет, просторова статистика

# UNDERSTANDING STATISTICAL DATA FOR MAPPING PURPOSES

***Aileen Buckley***

***Esri, Redlands, USA***
***E-mail:***abuckley@esri.com

Fundamentally, maps display only two types of data: qualitative and quantitative. Qualitative data differentiates between various types of things. Quantitative data communicates a message of magnitude.

A number of mapping methods have been developed that combine various map features and symbols. Choropleth mapping uses lightness to symbolize polygons. Proportional symbol maps display results as points that vary in size based on their associated values.

Because most statistical data is quantitative in nature, this article focuses on mapping quantitative data. However, to appropriately map quantitative data, you must understand it. Not all methods work equally well for all quantitative data.

Counts and characteristics can be used to derive measures that express either summarizations (e.g., mean, median) or relationships (e.g., densities, proportions).Tabulations and derived values for enumeration units are assumed to be uniform across the area and change at unit boundaries (i.e., they do not blend from one unit into another).

You must also consider whether the statistic being mapped depends on the size of the unit. Counts or totals and measures, such as area and perimeter, are summary statistics for the unit and are only true when they represent the unit as a whole. These statistics are said to be spatially extensive. The statistic is the sum of the properties of elements that make up the unit.

In contrast, values such as population density or cancer rates can describe any part of the unit (if the unit is assumed to be homogeneous). These statistics are *spatially intensive* and do not depend on the size of the unit. If you divide the unit, the value will stay the same. However, values for spatially extensive data cannot stay the same.

Spatially intensive data can be derived from spatially extensive data. For example,dividing counts by area yields density or dividing the count for one unit by the sum of counts for all units yields a proportion.

Counts or totals and other spatially extensive data should never be symbolized using the choropleth mapping method.

Why? Because this method does not accurately represent the nature of the data. Mapping spatially extensive data using a choropleth method masks the concentration of features within the areas because it assumes the distribution is uniform.

The choropleth method also masks distributions that are uniform. Different units on the map cannot be compared because no consistent denominator has been used to provide a basis for comparison.

To correct the problems caused by mapping counts using the choropleth method, you can convert the data to the correct type so it can be shown by lightness within areas. This is often necessary for data represented as points, lines, or rasters and with other mapping methods such as proportional circle.

Do this by normalizing or standardizing the data. These two terms, often used interchangeably, are slightly different.

Normalizing the data scales all numerical values to a range from zero to one.Standardization transforms the data so that is has zero mean and unit variance. Both techniques have drawbacks. If the dataset has outliers, normalizing will scale the normal data to a very small interval. When using standardization, the assumption is that the data has been generated with a certain mean and standard deviation, although this may not be the case.

In mapping, cartographers often use the term *derived data* to refer to data that has been transformed through normalization or standardization so it can be compared in a meaningful way. Transformations commonly used in mapping include ratios or rates, proportions, percentages, and densities.

It is important to differentiate between spatially intensive and spatially extensive measures. Density is a spatially extensive measure. A proportion, generated by dividing the number of items in a unit by the total number of items, is spatially extensive because the number per unit has been divided by a constant (the total number of things). For derived values such as proportions,percentages and rates, the resulting numbers can only be true for the *entire* unit, not parts of it. For units of intrinsic importance (e.g., counties) mapping the proportion of the value allocated to eachunit should not be mapped using the choropleth method. In such cases, it may be best to use graduated symbols.

Understanding more about the nature of the statistical data used for mapping purposes will help you better understand the methods that can be used to map it.Ultimately, the goal is to match appropriate data with the most effective method so that your map can be easily, quickly, and correctly interpreted by your readers.

***Keywords:*** qualitative, quantitative, spatially intensive, spatially extensive, choropleth method, spatial statistics.

**References**

1. Brewer, Cynthia A. 2006. "Basic Mapping Principles for Visualizing Cancer Data Using Geographic Information Systems (GIS)," www.ajpmonline.org/article/S0749-3797%2805%2900358-2/fulltext.
2. Cote, Paul. Effective Cartography: Mapping with Quantitative Data, www.gsd.harvard.edu/gis/manual/normalize/.
3. Kimerling, A. Jon, Aileen R. Buckley, Phillip C. Muehrcke, and Juliana O. Muehrcke. 2011. Map Use: Reading, Analysis, Interpretation, Seventh Edition. Redlands, CA: Esri Press, 581 pages.
4. Longley, Paul A., Michael F. Goodchild, David J. Maguire, and David W. Rhind. 2011. Geographic Information Systems and Science, Third Edition, New York: Wiley, Chapter 4.
5. Pitzl, Gerald. 2004. Encyclopedia of Human Geography. "Choropleth maps."
6. Robinson, Arthur H., Joel L. Morrison, Phillip C. Muehrcke, A. Jon Kimerling, and Stephen C. Guptill. 1995. Elements of Cartography, Fifth Edition. New York: John Wiley & Sons, Inc., 674 p.
7. Saitta, Sandra. "Standardization vs. normalization," Data Mining Research blog.

*Поступила в редакцию 18.04.2013 г.*