Predictive Analytics

Geo Analytics – Part II

Prof. Dr. Volker Herbort

Technik
Informatik & Medien
Hochschule Ulm
University of
Applied Sciences

Contents

Hochschule Ulm
University of Applied Sciences

- Introduction
- Geospatial Data
- **Mapping Data**
- Points
- Geo Analytics ML

Mapping Data

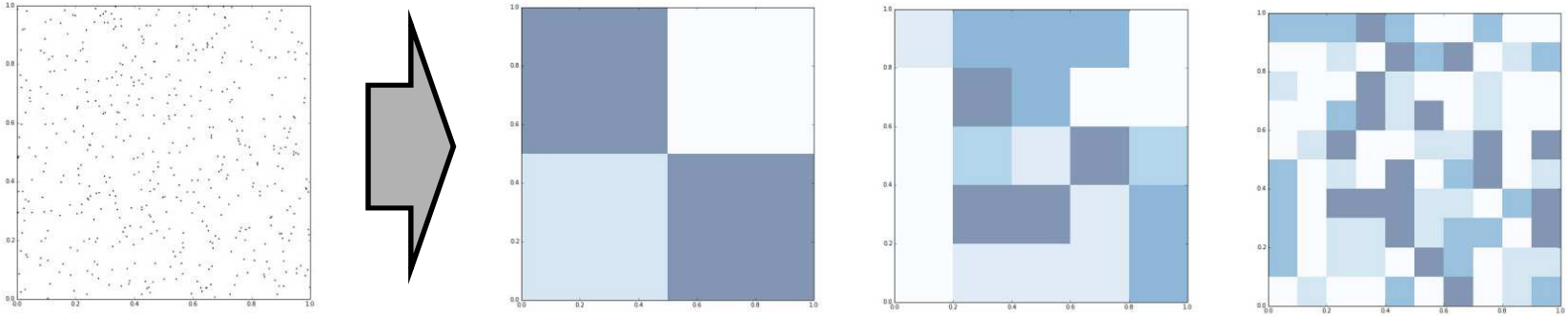
Hochschule Ulm
University of Applied Sciences

- Data maps
  - Abstraction from the purely geographical map
  - Representing numerical values within a spatial context
- Mapping data
  - A geographical approach to statistical visualization
  - The spread of data is considered in its geographical dimension

Modifiable Areal Unit Problem

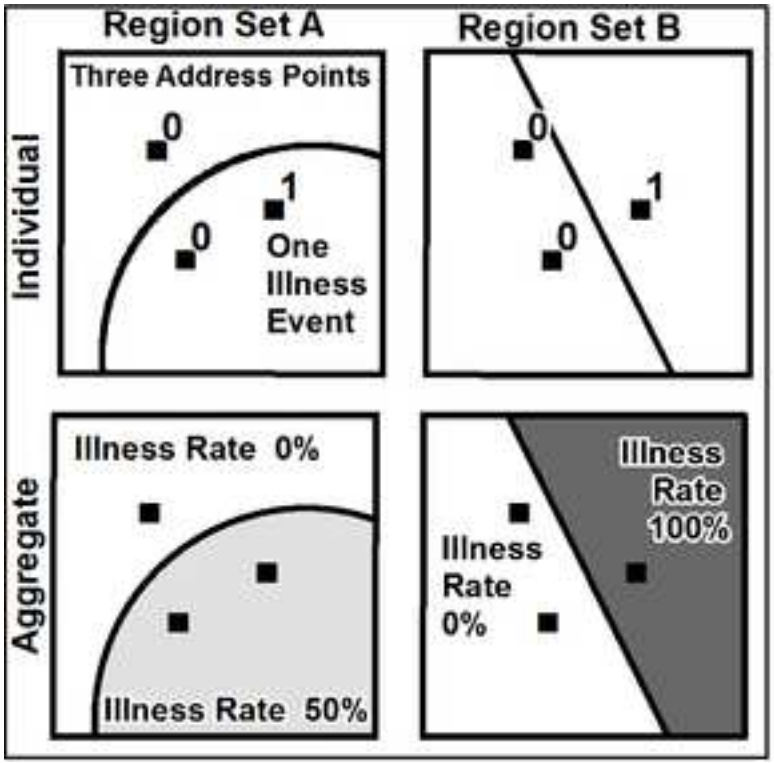
Hochschule Ulm
University of Applied Sciences

- Scale and delineation mismatch between:
  - Underlying process (e.g. individuals, firms, shops)
  - Unit of measurement (e.g. neighborhoods, regions, etc.)



Source: [2]

MAUP Example



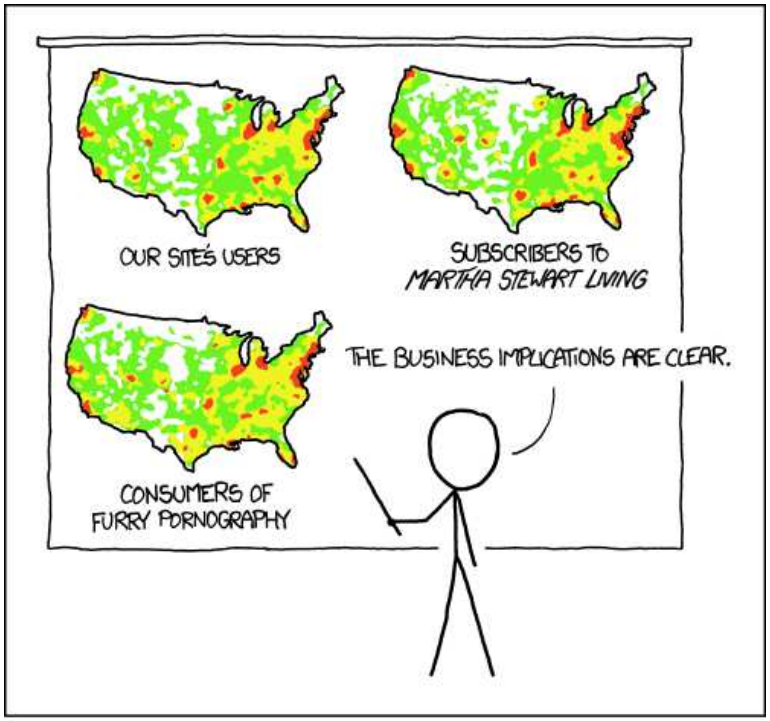
Source: [8]

- Always keep **MAUP** in mind when exploring aggregated data!!!

Choropleths

- Thematic map in which values of a variable are encoded using a color gradient of some sort
  - Counterpart of the histogram
  - Values are classified into specific colors
    - Colors --> in alignment with the goal of the map
    - Bins --> How many?
    - Algorithms:
      - Unique values
      - Equal interval
      - Qua/Quintiles (equal count)
      - Fisher-Jenks
  - Information loss as a trade off for simplicity

Beware standardization!!!

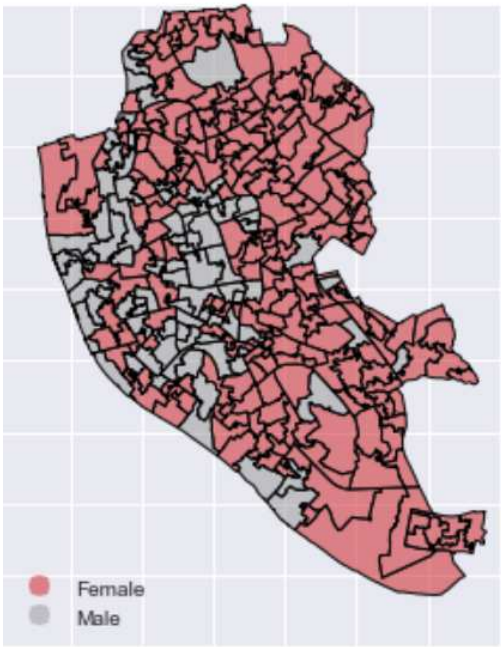


PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

Source: [5]

Unique values

- Categorical data
  - No gradient
    - Highlight differences but no distance
    - Use of appropriate color scheme!!!
  - Examples
    - Religion
    - Country of origin
    - Gender
    - …

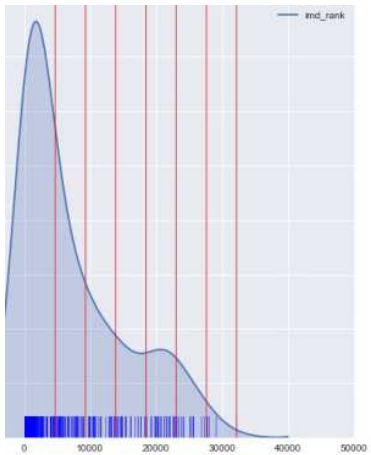


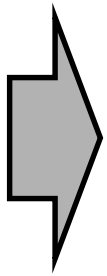
Source: [2]

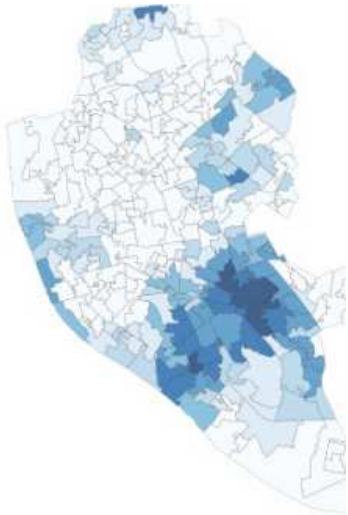
Equal interval

- Numerical Data
  - Split the value span of the data equally
    - Splitting based on the numerical value in question
  - Attention: Gives more weight to outliers

if the distribution is skewed

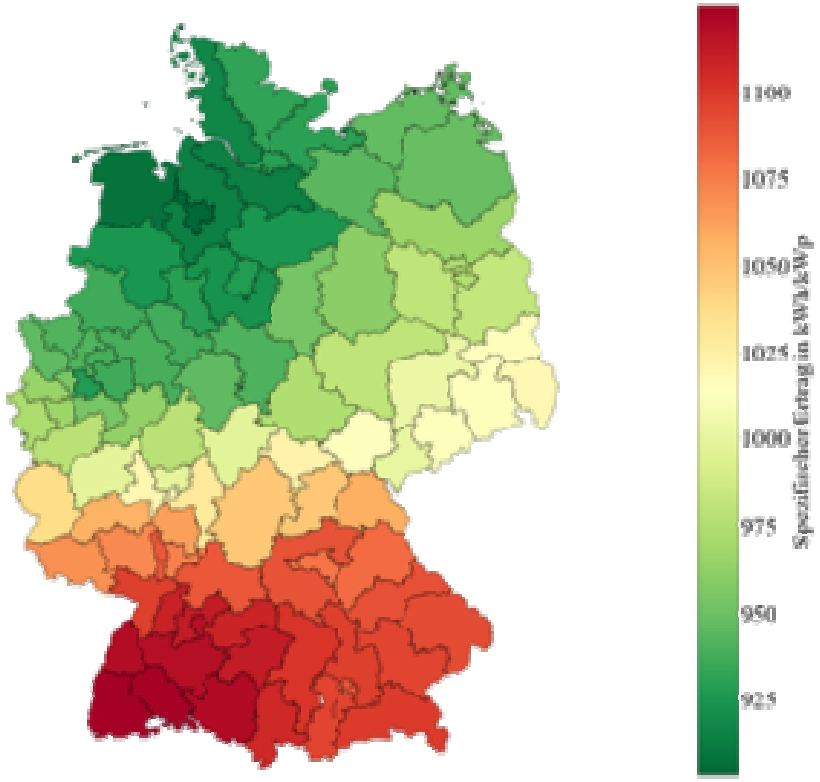






Source: [2]

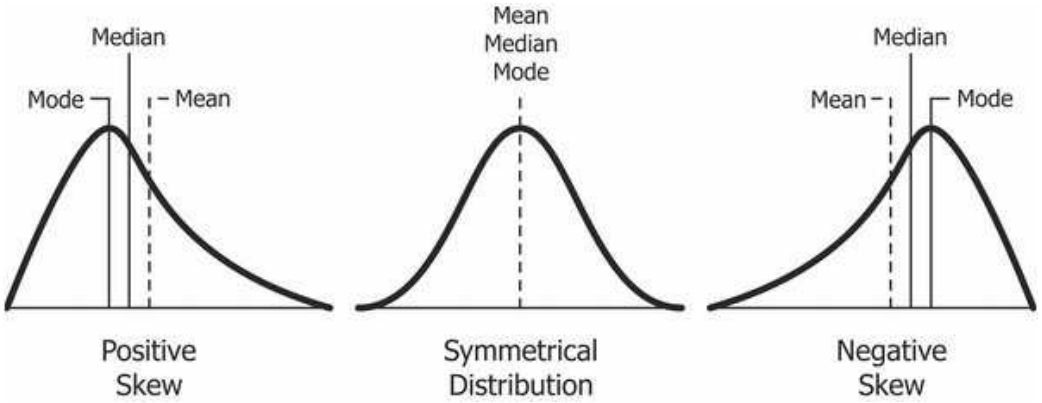
Equal Interval ctd.



Source: [6]

Quantiles

- Split the distribution keeping the same amount of values in each bin
  - Regardless of numerical values
  - Splitting based on the rank of the value
- Attention: If distribution is skewed, it can put very different values in the same bin



Source: [7]

Advice

- Different classification schemes can produce widely different maps as a result of:
  - The distribution of the values
  - The inherent simplification that a choropleth implies
- Explore different ones and combine choropleths with other graphical devices like histograms or density plots => Extensive data profiling

Cartograms

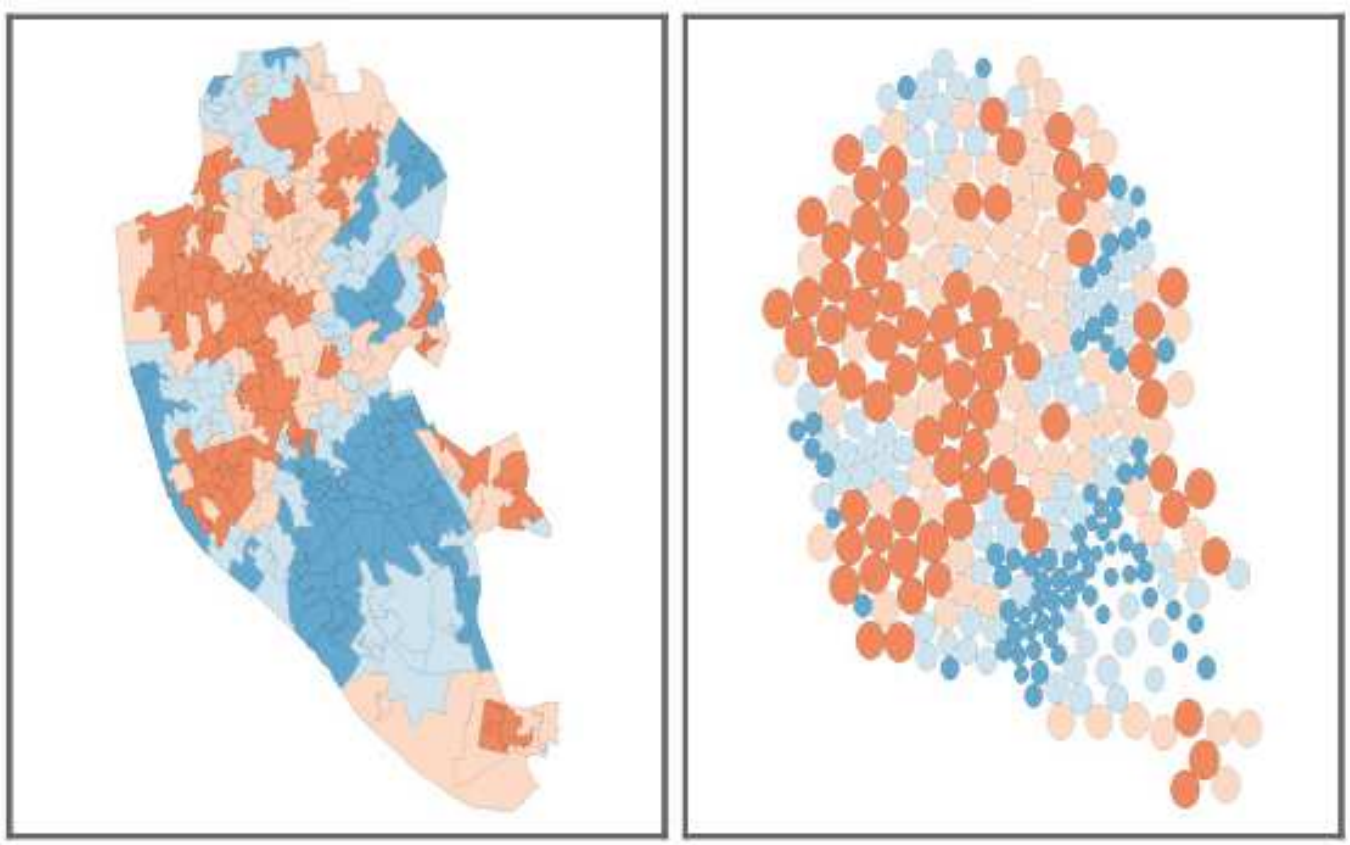
Hochschule Ulm
University of Applied Sciences

- Data maps where the variable is encoded, not by a color gradient, but by distorting the shape/size of the geographical objects
  - Useful in cases where the natural size/shape induces to wrong interpretation or obscures the intended representation.
  - If not done carefully, it can distort the message in unintended ways

Cartograms ctd.

Hochschule Ulm
University of Applied Sciences

- Shape size and value missmatch



Source: [2]

Visual Data Design

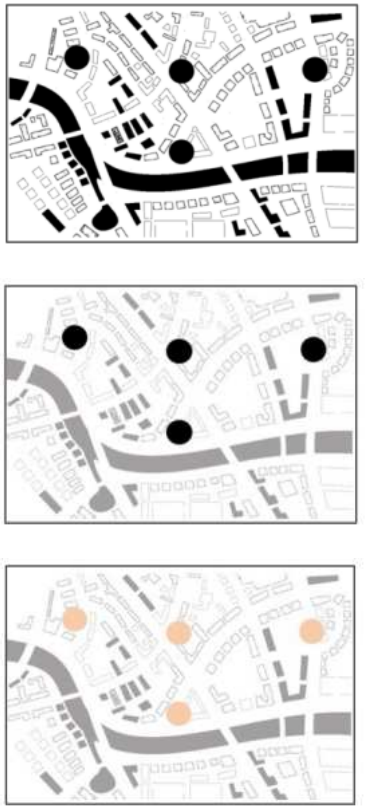
Hochschule Ulm
University of Applied Sciences

- Design principles for good maps
  - Always display legends
  - Remove elements without meaning
  - Avoid radiant colors
  - Remove irrelvant lines and frames

Layering

Hochschule Ulm
University of Applied Sciences

- Use of colours for layers
  - Same colour scheme as background hides objects
  - Highlight objects on layer
    - Saturation of colours
    - Different colours

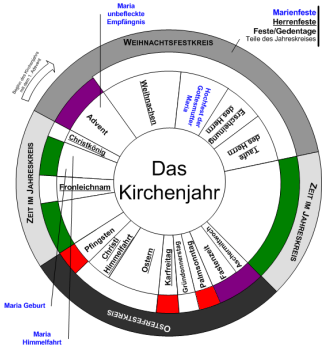


Source: [8]

Colours

- Colours come with associations
  - Feelings
  - Conditions
  - Marketing
  - Religion

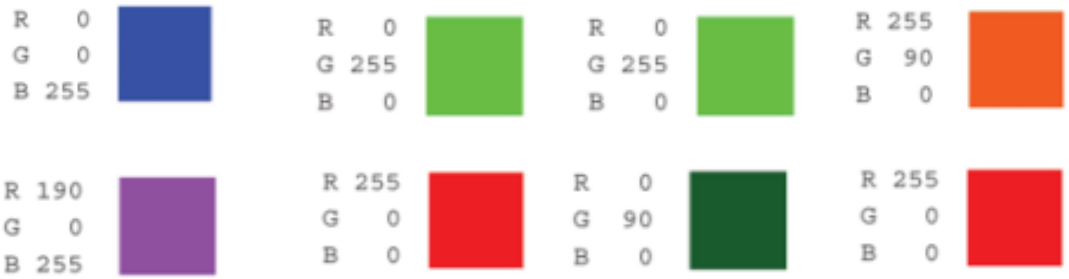
Source: [8]



Colours ctd.

Hochschule Ulm
University of Applied Sciences

- Colours have different contrast values
  - Visual disability access
  - Appealing colours can cause problems
  - Certain combination should be avoided



Source: [8]

Colours ctd.

Hochschule Ulm
University of Applied Sciences

- Use different colour palettes for different data
  - Categorical
    - No order in values
    - Colours used to distinguish

Source: [8]

  - Ordinal
    - Colours should indicate distance in value
      - sequential palette
      - diverging palette

Source: [8]

Contents

Hochschule Ulm
University of Applied Sciences

- Introduction
- Geospatial Data
- Mapping Data
- **Points**
- Geo Analytics ML

Points

Hochschule Ulm
University of Applied Sciences

- Points **like** polygons
  - Points can represent "fixed" entities
  - Points are qualitatively similar to polygons/lines
  - The goal here is, taking location fixed, to model other aspects of the data
  - Examples
    - Cities (in most cases)
    - Buildings
    - Polygons represented as their centroid

Points

Hochschule Ulm
University of Applied Sciences

- Points **unlike** polygons
  - Points can also represent a fundamentally different way to approach spatial analysis
    - Rather than exhausting the entire space, points can be events subject to occur anywhere
    - The location of the event is part of what we are trying to understand/model
    - The interest focuses on characterizing the pattern that the points follow over space

Example: Geo-tagged tweets

Hochschule Ulm
University of Applied Sciences

Source: [2]

Points Patterns

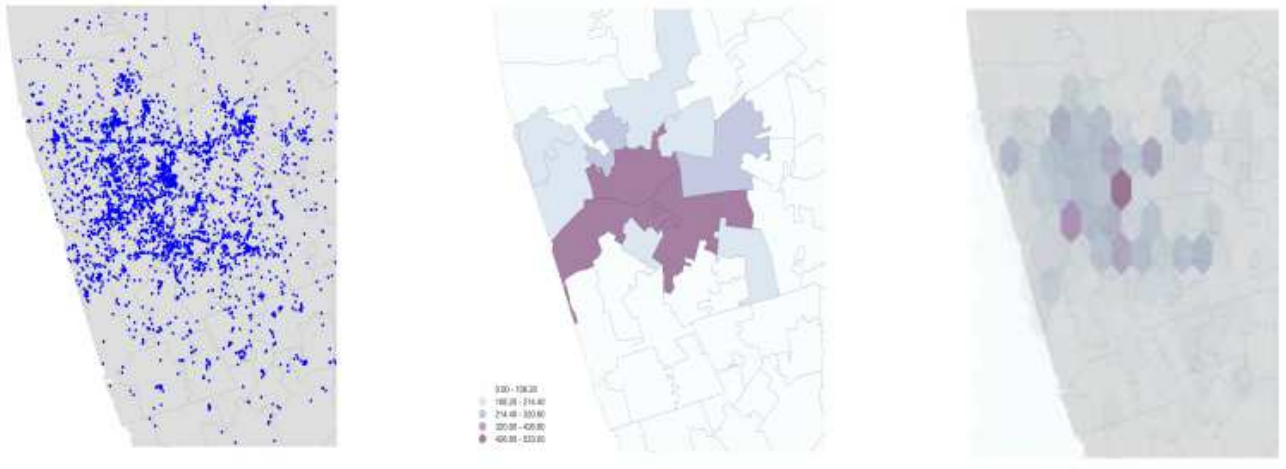
- Distribution of points over a portion of space
  - Assumption is a point can happen anywhere on that space, but only happens in specific locations
    - Unmarked: locations only
    - Marked: values attached to each point
  - Point Pattern Analysis
    - Describe, characterize and explain point patterns, focusing on their generating process
      - Visual exploration
      - Clustering properties
      - Statistical modeling of the underlying processes

Points meet polygons

Hochschule Ulm
University of Applied Sciences

- Use polygon boundaries and count points per area
  - Create choropleth mapping (Beware of MAUP)
  - Polygons need to "make sense"

  => delineation relates to the point generating process



Source: [2]

Kernel Density Estimation (KDE)

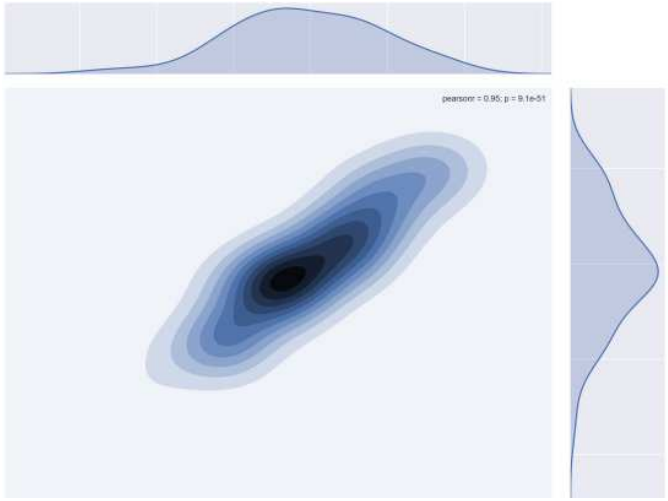
Hochschule Ulm
University of Applied Sciences

- Estimate the continous observed distribution of a variable
  - Probability of finding an observation at a given point
  - "Continuous histogram"
  - Solves (much of) the MAUP problem, but not the underlying population issue

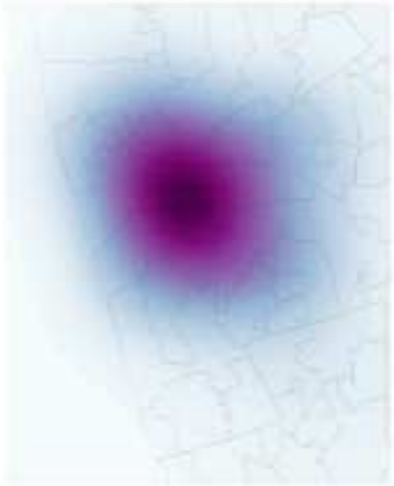
Bivariate (spatial) KDE

Hochschule Ulm
University of Applied Sciences

- Probability of finding observations at a given point in space
  - Bivariate version: distribution of pairs of values
  - In space: values are coordinates (XY), locations
  - Continuous "version" of a choropleth



Source: [2]



Source: [2]