M.C. Thrun, A. Ultsch

Effects of the pay-out system of income taxes to municipalities in Germany

Feel free to contact me through www.deepbionics.org

Main Problem

- Distribution analysis has the goal to estimate the distribution of a variable:
  - Comparison to standard distributions
  - Estimation of probability density function (pdf)

- Many approaches are used in Distribution Analysis
  - Every approach has a assumptions behind it, e.g.
    - Descriptions of distributions using a single distribution, like Lognormal or Gamma are often quite weak in describing the tails of the distribution [Dagum, 1977]

      -> Often separate models for the upper vs. lower parts of distributions [Richmond, Hutzler, Coelho, & Repetowicz, 2006, p. 140]
    - Sometimes better: Gaussian mixture models [Thrun & Ultsch, 2015]

- Knowledge Discovery using low-level methods possible, e.g.
  - Income tax system of German municipalities

Motivation

- Applications of Distribution Analysis
  - Detect meaningful structures
  - Carefully preprocess variables for cluster analysis or supervised machine learning, e.g.
    - Choice of an appropiate transformation for each variable
  - Descriptive statistics, e.g. estimation of variance
  - Selection of correct statistical test
    - Check conditions

- Solution
  - Combine different approaches

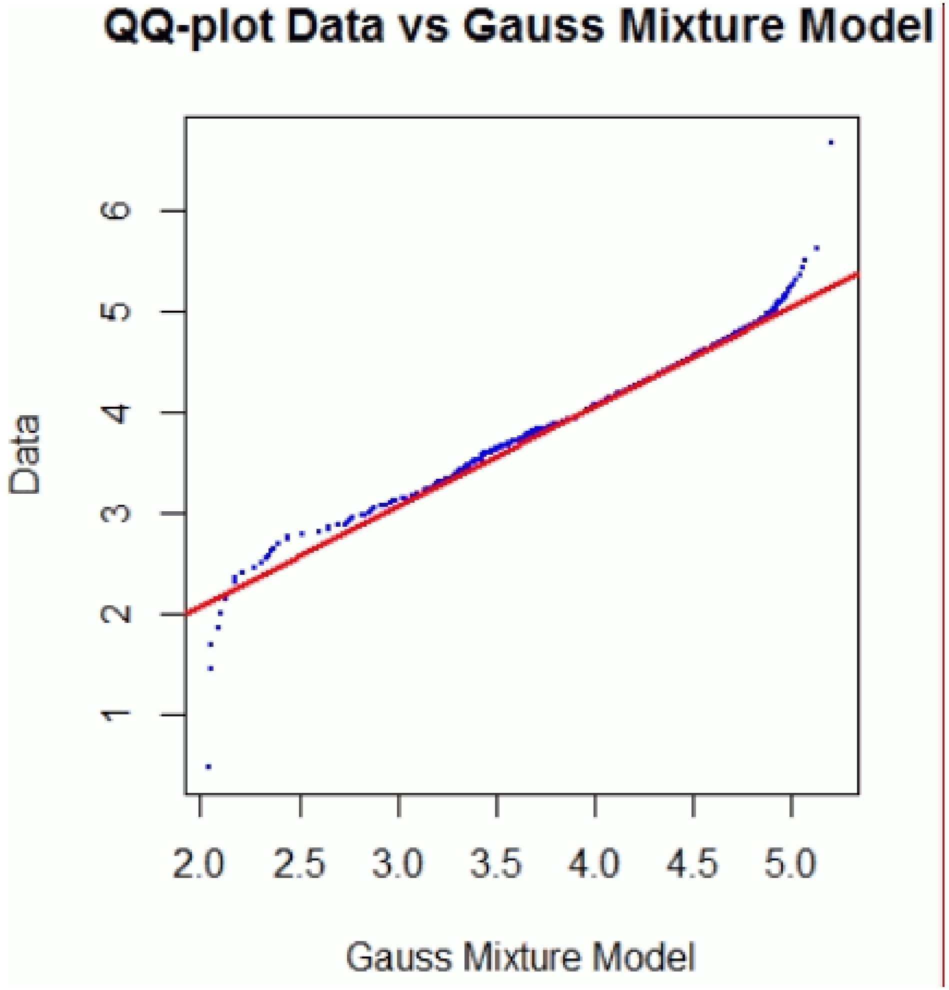
Most conventional approaches

- Histogram
  - Kernel density estimation with fixed radius
    - The choice of the width and number of bins is critical for the right fit of the probability density function
  - Use Optimal bin width under the assumption that the variable is Gaussian distributed [Keating & Scott, 1999]

- The Box-Whisker diagram (box plot)
  - Visualizes the number of values in a specific range
    - End of the two whiskers are proportional to the interquartile range (often 1.5*IQR), [Tukey, 1977]
    - The box marks 25 and 75% percentile
  - Does not indicate multimodality or if median is valid
    - We propose to use PDE optimized violin plots for multimodal distributions

Distribution Analysis: QQ plot

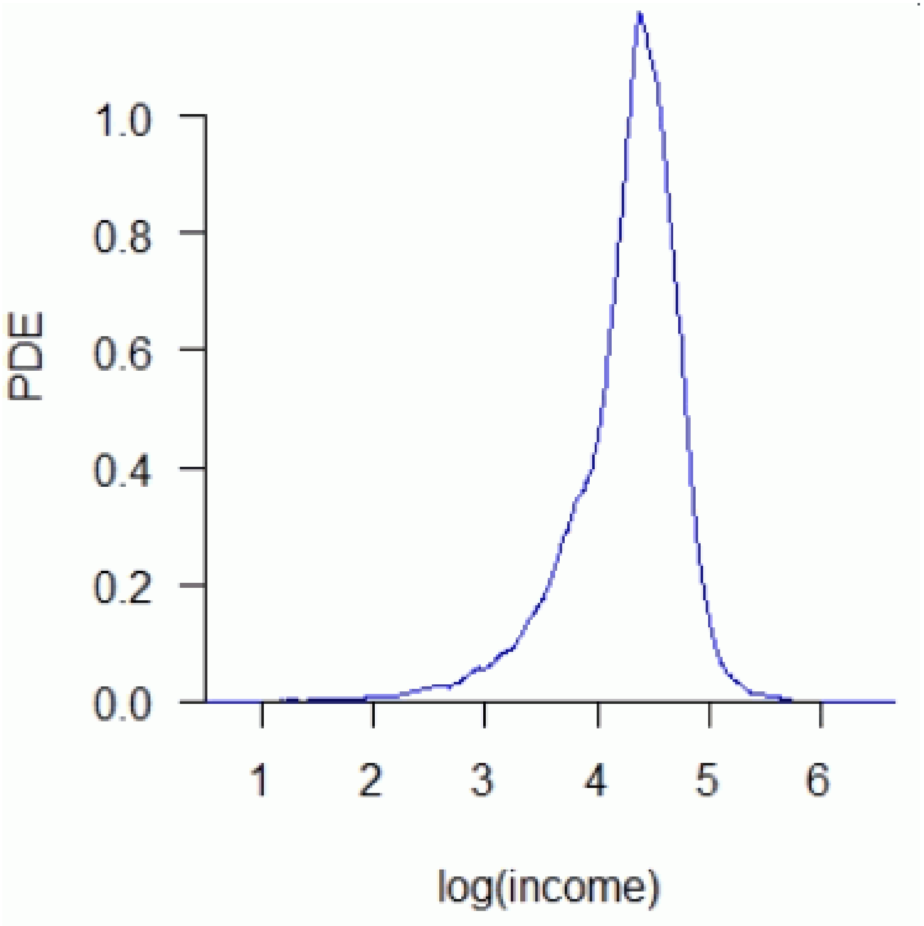
- Another good way to explore the distribution of a feature is a comparison with a known distribution
- QQ plot [Michael 1983]
  - Compares two distributions by using n quantiles
  - Empirical distribution vs known distribution
  - If straight line: distributions equal
  - The Gaussian distribution is an ideal starting point for such a comparison
- In principle every common distribution should be checked



Distribution Analysis: PDE plot

- Pdf estimation called "Pareto Density Estimation (PDE)
- Kernel density estimation with variable radius
  - Representing the relative likelihood of a given variable taking on specific values
  - Slivered in kernels with a specific width
    - this width, and therefore the number of kernels, depends on the data
  - Particularly suitable for the discovery of structures in continuous data
  - Allows the discovery of mixtures of Gaussians (Ultsch, 2005a)

-> PDE is designed in particular to identify groups in data [Ultsch 2005]



Example: German Tax System

- Several layers of administration and legislation are involved
  - Hinders an easy comprehension of the system
- Spatial Unit: municipality:
  - National legislation demands of the system that
    - Output should be a fixed proportion of input
  - Input: total income tax yield of each municipality
  - Output: share of income tax revenues
    - the funding a municipality receives from the state

Features: MTY and ITS in 2010

- Dataset of [Ultsch & Behnisch, 2017] for the year 2010
  - 11,669 municipalities
  - 228 so-called “unincorporated areas” generally forested areas, lakes and larger rivers
- The number of taxpayers per municipality was given
- Municipality Income Tax Yield (**MTY) – Input**
  - Data not directly available

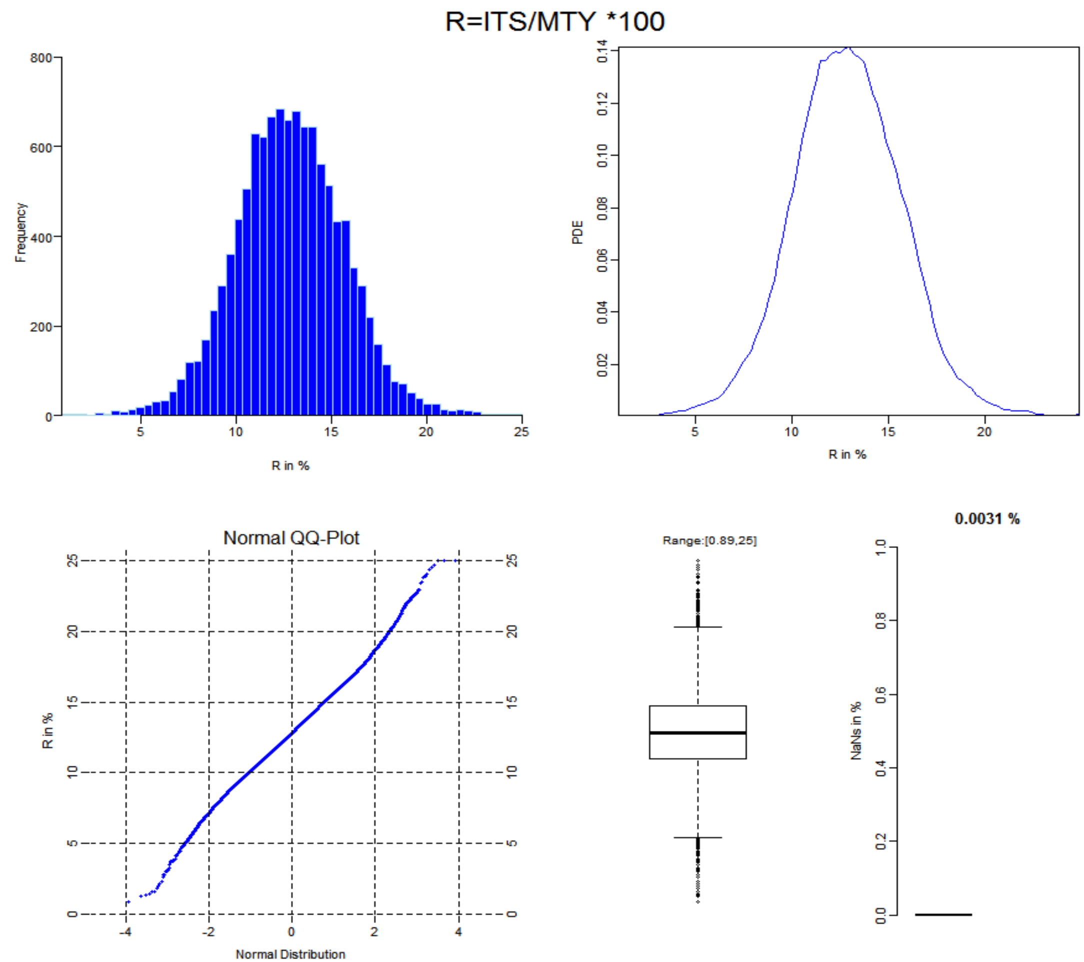
  -> Estimate MTY
    - Taxes per payer of 2007 and 2010 of the Regional Database Germany [Destatis, 2015] was used
- The income tax share (**ITS**) of a municipality - **Output**
  - Sum of the income tax revenues paid by the state to a municipality was obtained from Regional Database Germany [Destatis, 2015]
    - Tax share divided by the number of taxpayers

Prior Model Assumption

- Expected funding of the state defined by ratio R

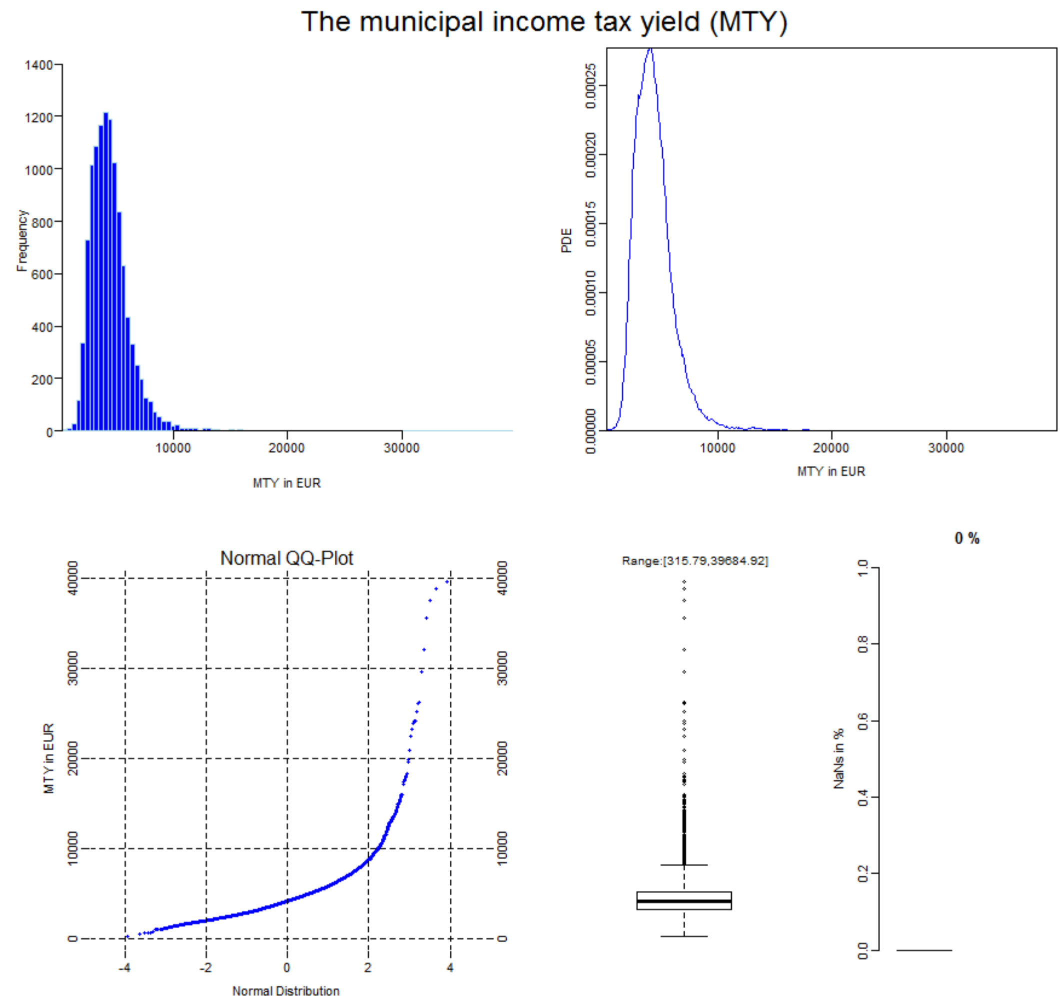
R = ITS/MTY*100

- The ratio has a mean of 13% ± 3%
  - Expected amount of collected tax
- Gaussian Distributed
  - Implies that there should be only unintentional deviations from this general percentage.



Municipality Tax Yield (MTY)

- All approaches agree
  - Unimodal distributed
- Not Gaussian distributed
  -> QQ-plot
  - Left-Skewed
  - Right tailed
- Outliers in box plot

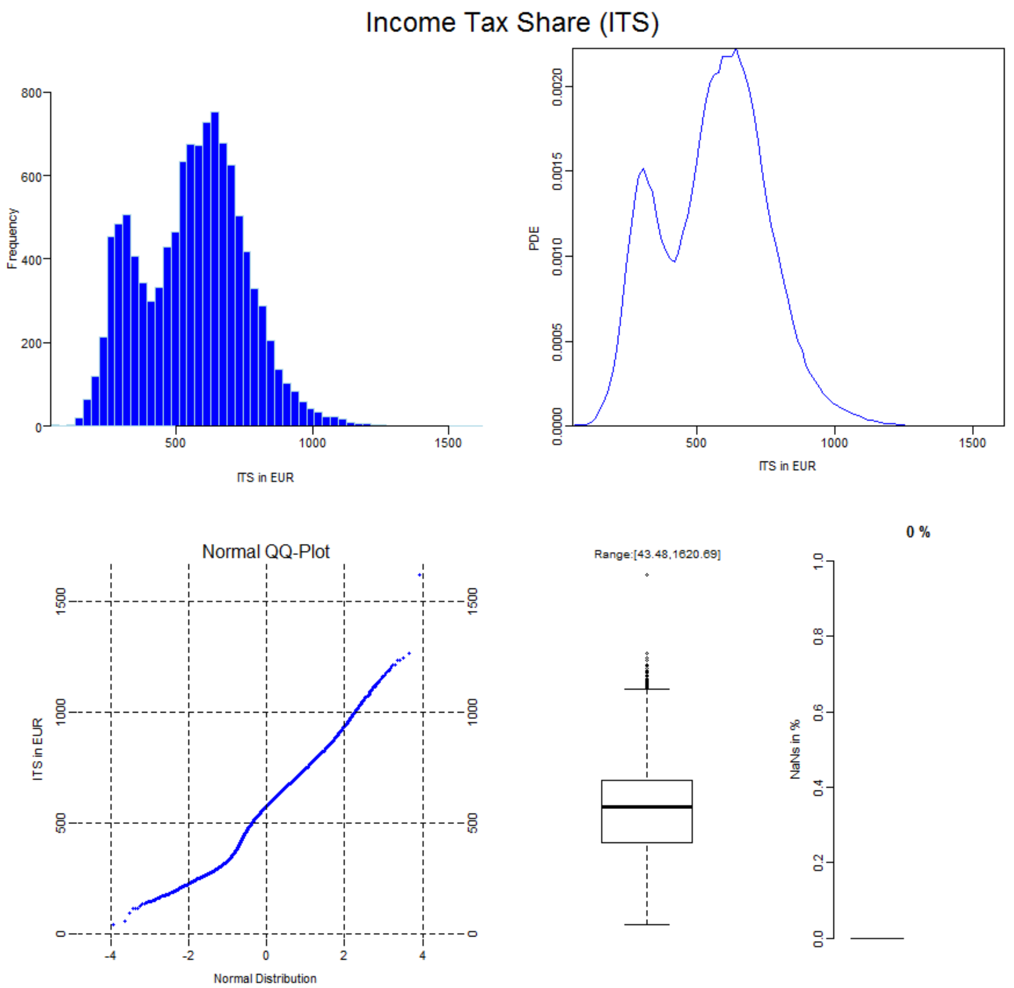


https://cran.r-project.org/web/packages/DataVisualizations/index.html

Income Tax Share (ITS)

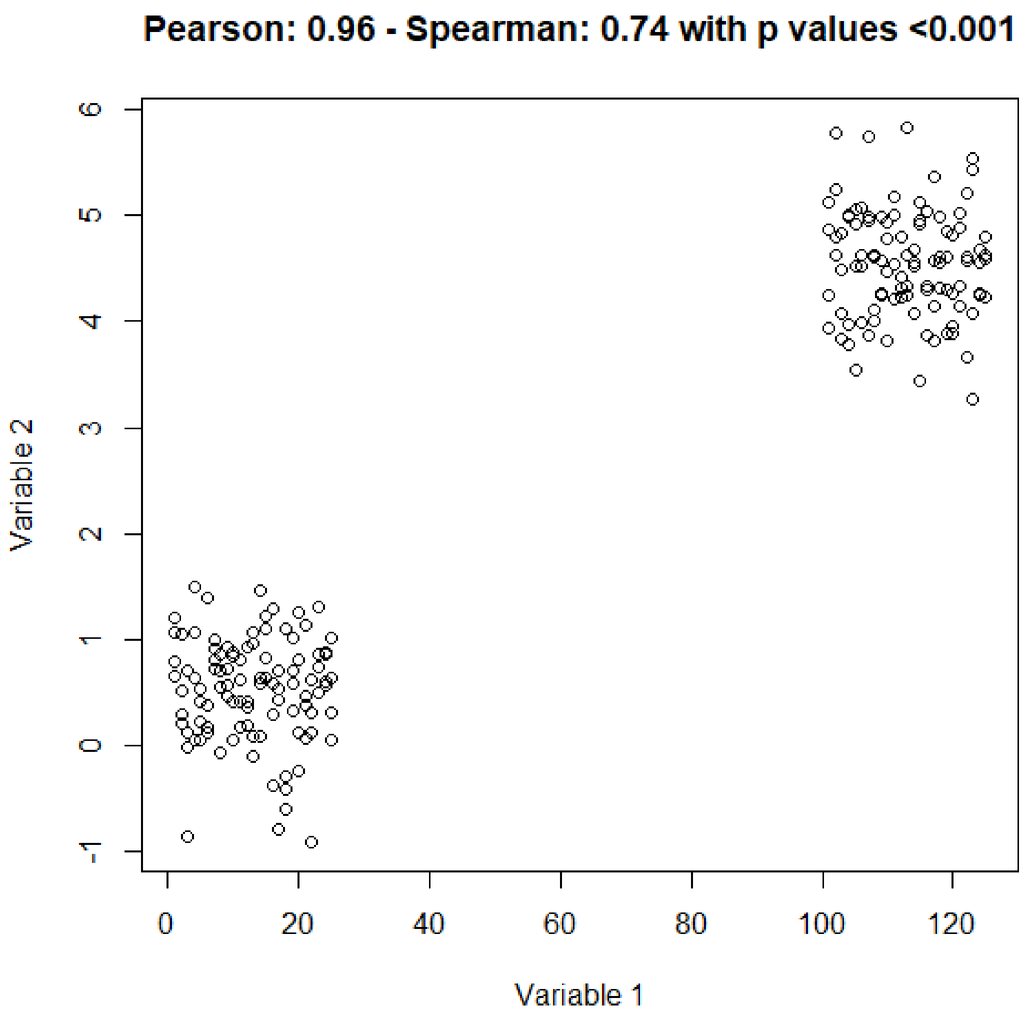
- Multimodal distributed
  - Two modes
  - One major outlier
- The first maximum of ITS lies at 300 EUR and the second at 640 EUR
- The deviations from the mean in ITS should unintentional

=> the distribution of ITS should be unimodal



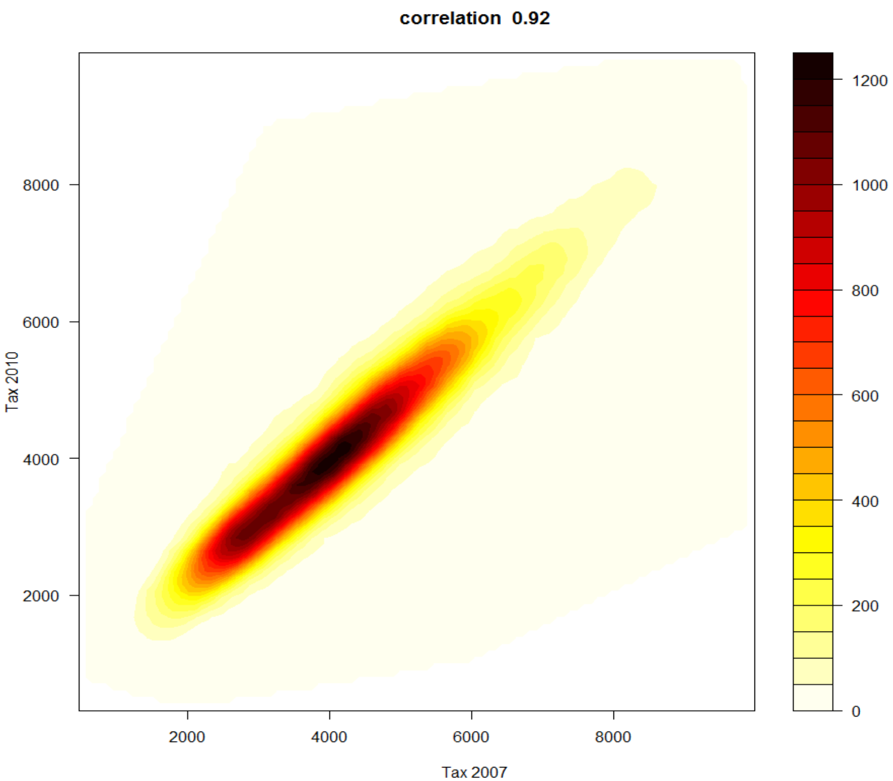
Relations between two Variables

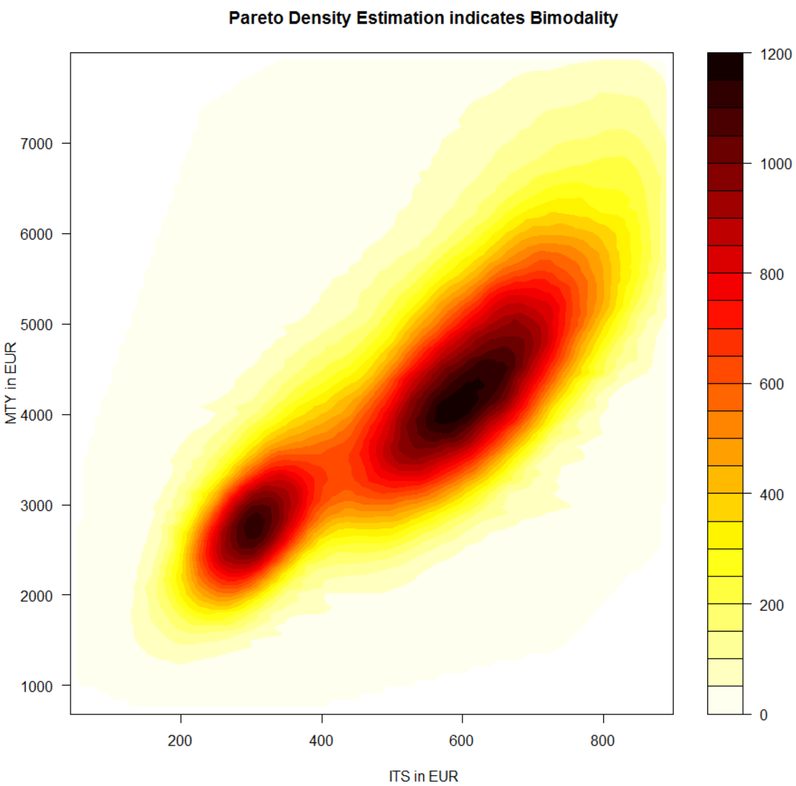
- Correlation values are often misleading
  - Use Scatteplots:
    - Two features are plotted against each other as points
  - Do not show mixtures due to overlapping of points
- Even Better: Two-Dimensional Density visualization
  - Use scatter density plot the densities
    - Estimation by PDE



Scatter Density plots using PDE

- MTY estimated through a scatter density plot between tax per payer in 2007 and Tax2010 (Left)
  - Shows a clear correlation with one mode
- Scatter density plot between the input MTY and the output (ITS) (Right)
  - Two modes are visualized.

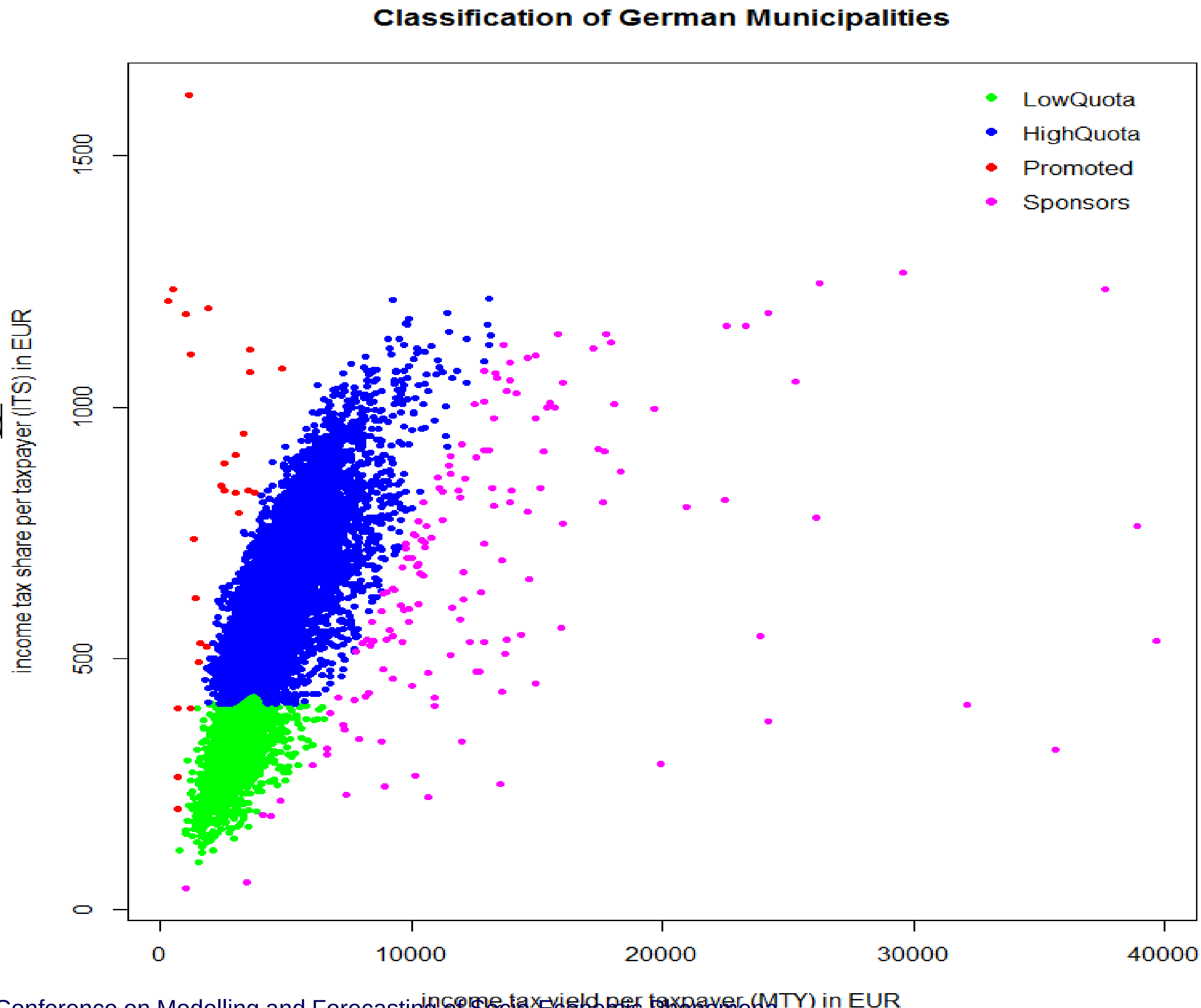




https://cran.r-project.org/web/packages/DataVisualizations/index.html

Two-Dimensional Bayesian classification

- Two-dimensional GMM is modeled
- Bayesian classification based on GMM calculated
- Points are colored by classification
- Scatter plot does not show the two modes



Knowledge Discovery

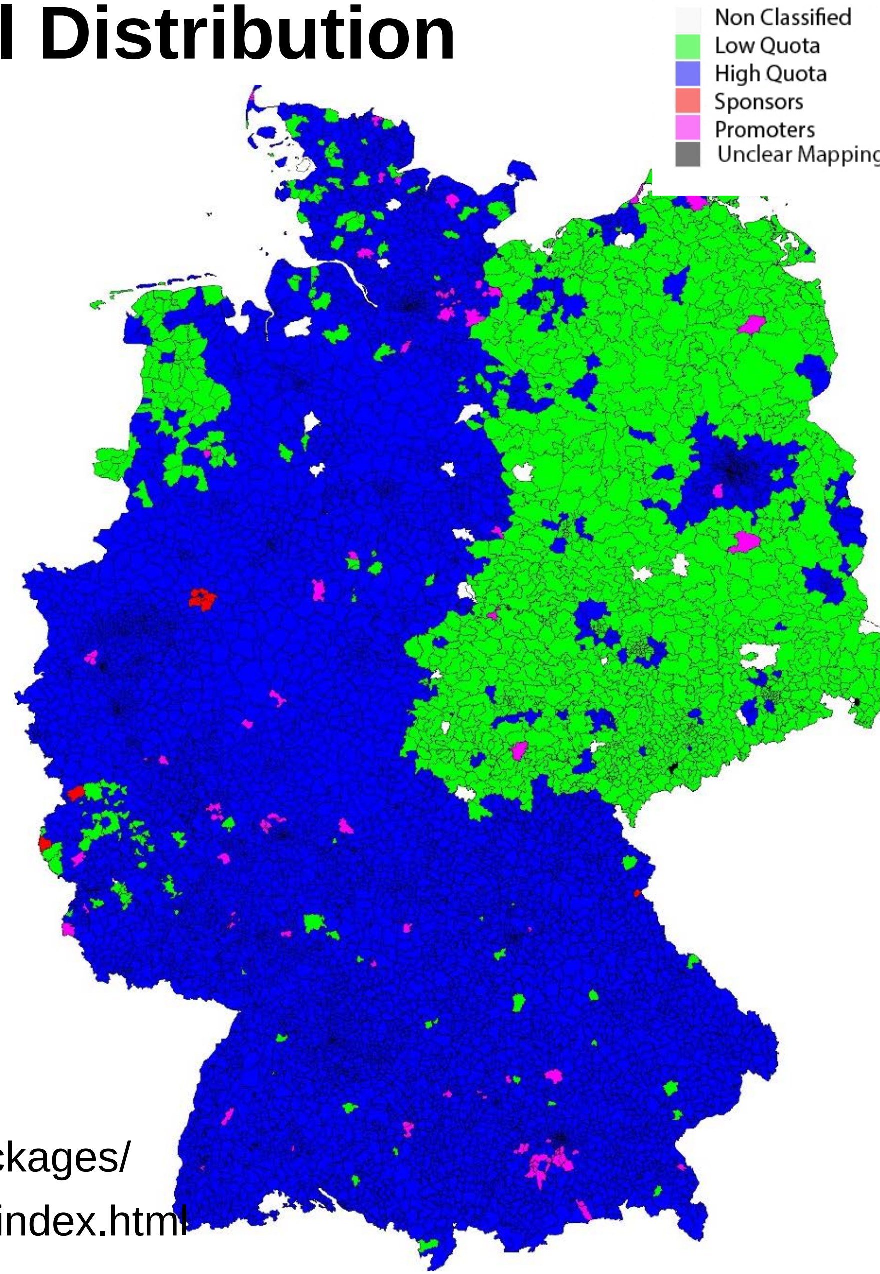
- Input of the system (MTY) was related to the output of the system (ITS)
  - Clear separation into two distinct distributions
  - Paying income tax per taxpayer of approx. 2,500 to 4,500 EUR to the state
    - the refund of a municipality can be either low or high
- Two-Dimensional GMM models the two states of the pay-out system of income taxes
  - Main classes - low quota vs. high quota
  - Outliers
    - Promoted class
      - municipalities receiving a substantially larger share of income taxes (30% and more)
    - Sponsors class
      - municipalities receiving a substantially smaller share of income taxes (8% and less)

Geographical Distribution

- In Germany every municipality has an Community Identification Number – „Amtlicher Gemeindeschlüssel“ (AGS)
- Shapefile of postal codes was available for Germany
- AGS were aggregated to postal codes (Mapping in package)
  - Colored by classification-

-> Political map is presented on the right

- For Poland with the right shapefile possible with our package

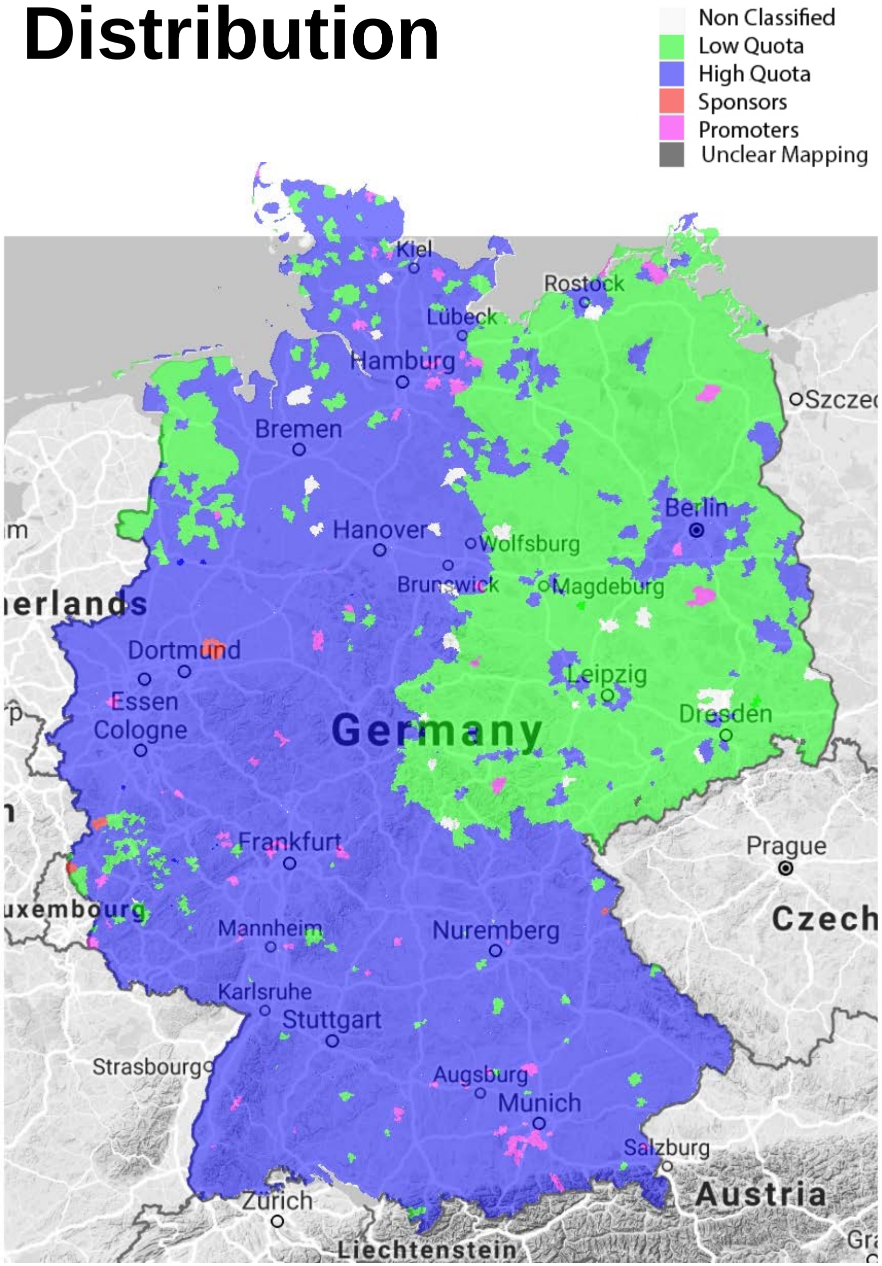


https://cran.r-project.org/web/packages/DataVisualizations/index.html

Geographical Distribution

- Choropleth map
  - thematic map with areas colored in proportion to the measurement
- Instead of statistical variable here colored by a classification
- Geographical distribution of the low quota vs. high quota municipalities reveals an evident east-west disparity

https://cran.r-project.org/web/packages/DataVisualizations/index.html



Discussion of Results

- Input and output of tax system should be proportional
  1. Income tax share per taxpayer (ITS) should be proportional to the municipal income tax yield (MTY).
  2. municipality should expect a certain fixed percentage of the taxes it delivers
- Detailed analysis of the distributions revealed
  - That observed probability distribution of ITS consisted of two distinct distributions.

  => Pay-out system of income taxes to municipalities operates in two distinct modes

      - Low quota vs high quota.

Summary

- Distribution Analysis itself allows to discover new knowledge
  - Methods were applied to Germany's complex system of allocating tax revenues
- Correct estimation of density is crucial for estimating the pdf and improves a scatterplot significantly
- Geographical distribution of the low quota vs. high quota municipalities revealed an evident east-west disparity
  - Percentage of income tax revenues a municipality received per taxpayer depended on the geographical location
  - If located in western Germany, the municipality could expect about 15-30%.
  - If located towards the eastern Germany, its share was more likely to be only 10% or less.

.

Thank you for listening.

Any questions?

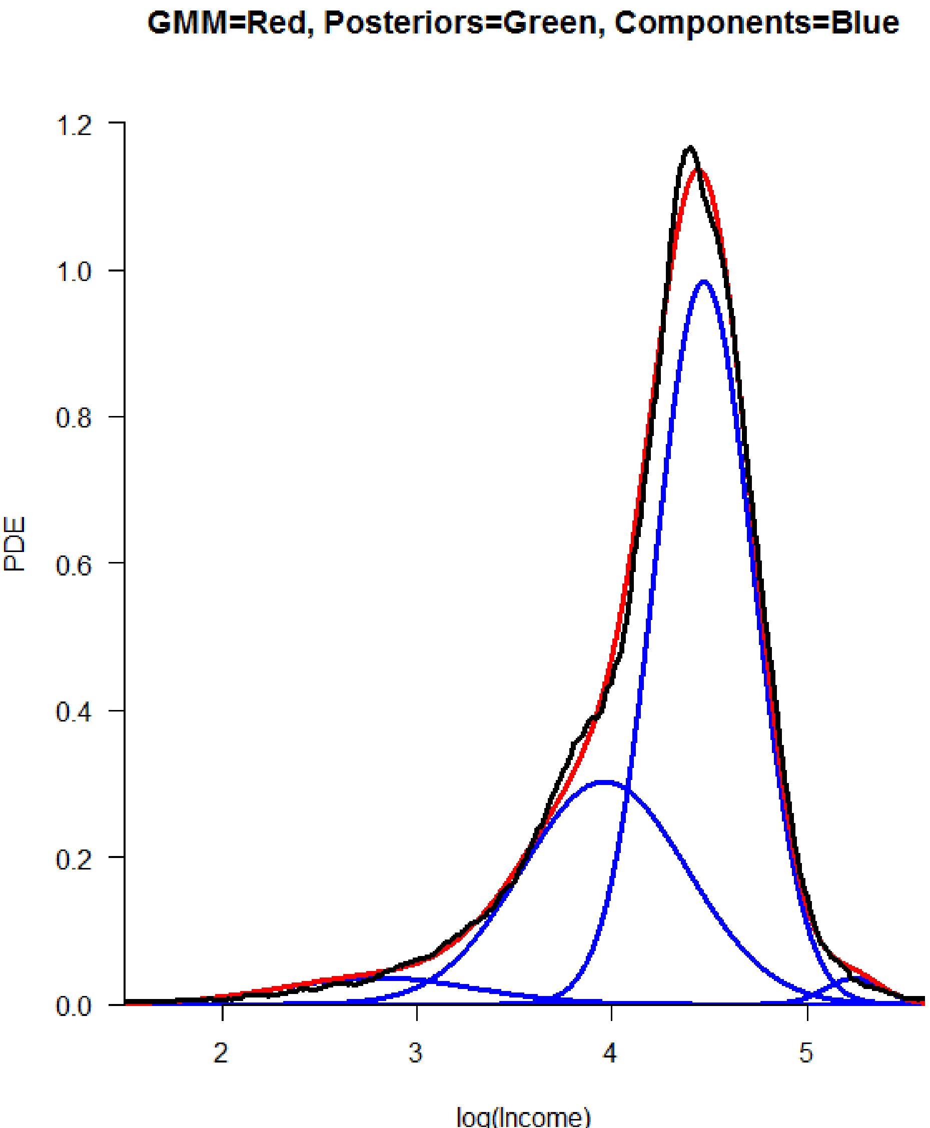
Example: One-dim. Gaussian Mixture Model (GMM)

Blue: Components $\mathrm{N}(m_i, SD_i)$

Red: $\mathrm{GMM}(x) = \sum_{\mathrm{i}=1}^{4} w_i * \mathrm{N}(m_i, SD_i)$

$$\sum_{i}^{4} w_i = 1 \qquad \int GMM\ (x) = 1$$

- Using Bayes-Theorem the EM-algorithm [Press 2007] estimates a log Gaussian mixture of four density states
- Through the likelihood to generate data in a component of the mixture p(x|c_i) we calculate the posterior p(c_i|x)



Thrun, M. C., & Ultsch, A. : Models of Income Distributions for Knowledge Discovery, Proc. European Conference on Data Analysis (ECDA), DOI: 10.13140/RG.2.1.4463.0244, pp. 136-137, Colchester, 2015.

Boundaries by using Bayes Theorem

Prior:

Probability to choose a class

conditional Probability: Likelihood to generate data in this class

$$p(c\_i \mid x) = \frac{p(x \mid c\_i)\, p(c\_i)}{\sum_{i=1}^{L} p(x \mid c\_i)\, p(c\_i)}$$

posterior

Normalization

$$\sum_{i=1}^{L} p(c\_i) = 1$$

$$\sum_{i=1}^{L} p(c\_i \mid x) = 1$$

Thrun, M. C., & Ultsch, A. : Models of Income Distributions for Knowledge Discovery, Proc. European Conference on Data Analysis (ECDA), DOI: 10.13140/RG.2.1.4463.0244, pp. 136-137, Colchester, 2015.

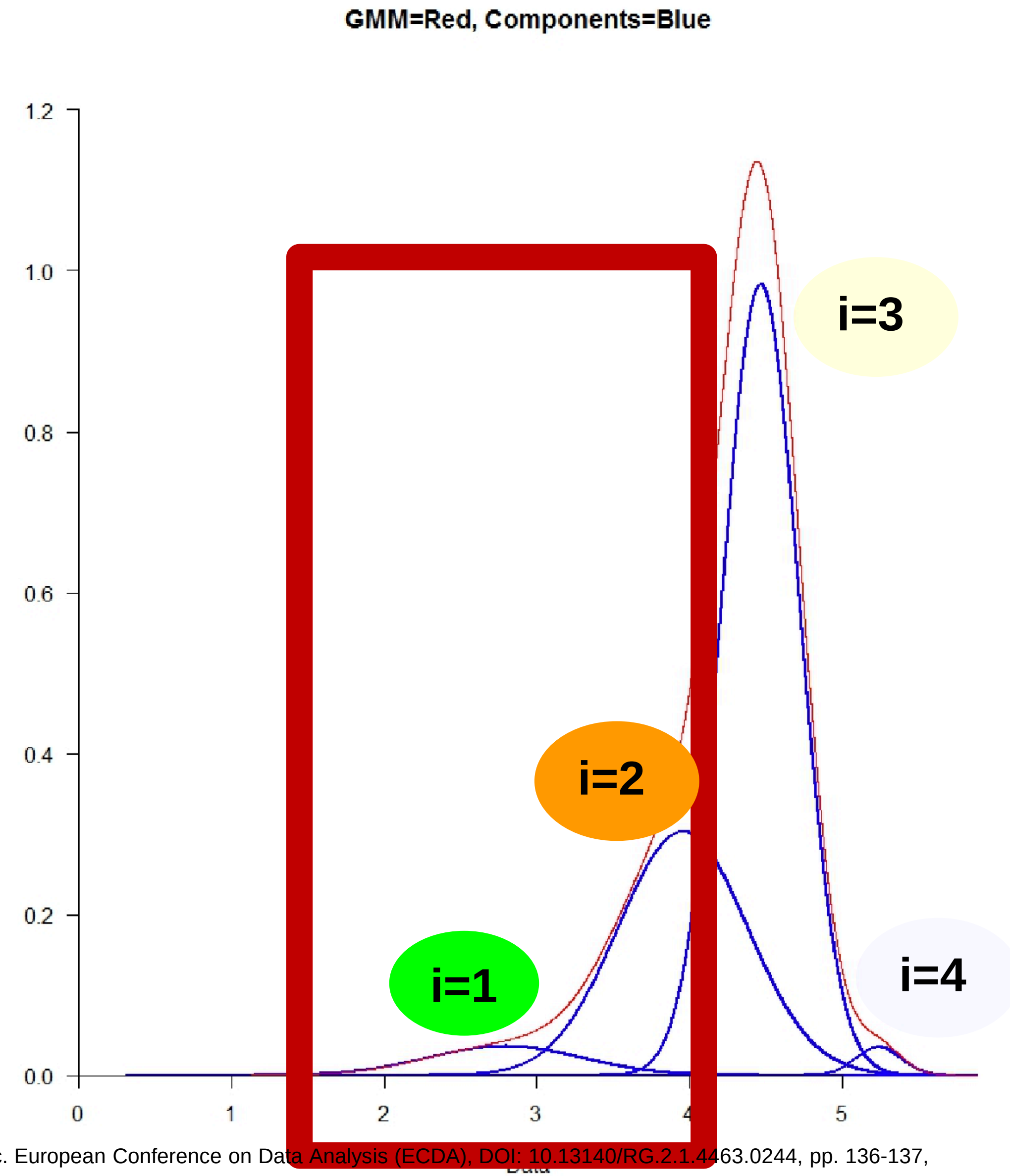
Application of Bayes Theorem

Blue: Components

Red: GMM(x)

Through the likelihood to generate data in a component of the mixture p(x|c_i) we calculate the posterior p(c_i|x)

- Example: Lets look at the red window with component 1 and component 2



Thrun, M. C., & Ultsch, A. : Models of Income Distributions for Knowledge Discovery, Proc. European Conference on Data Analysis (ECDA), DOI: 10.13140/RG.2.1.4463.0244, pp. 136-137, Colchester, 2015.

First Boundary in GMM

$$\mathrm{GMM}(x) = \sum_{i=1}^{4} \mathrm{w_i} * \mathrm{N}(\mathrm{m_i}, \mathrm{SD_i})$$

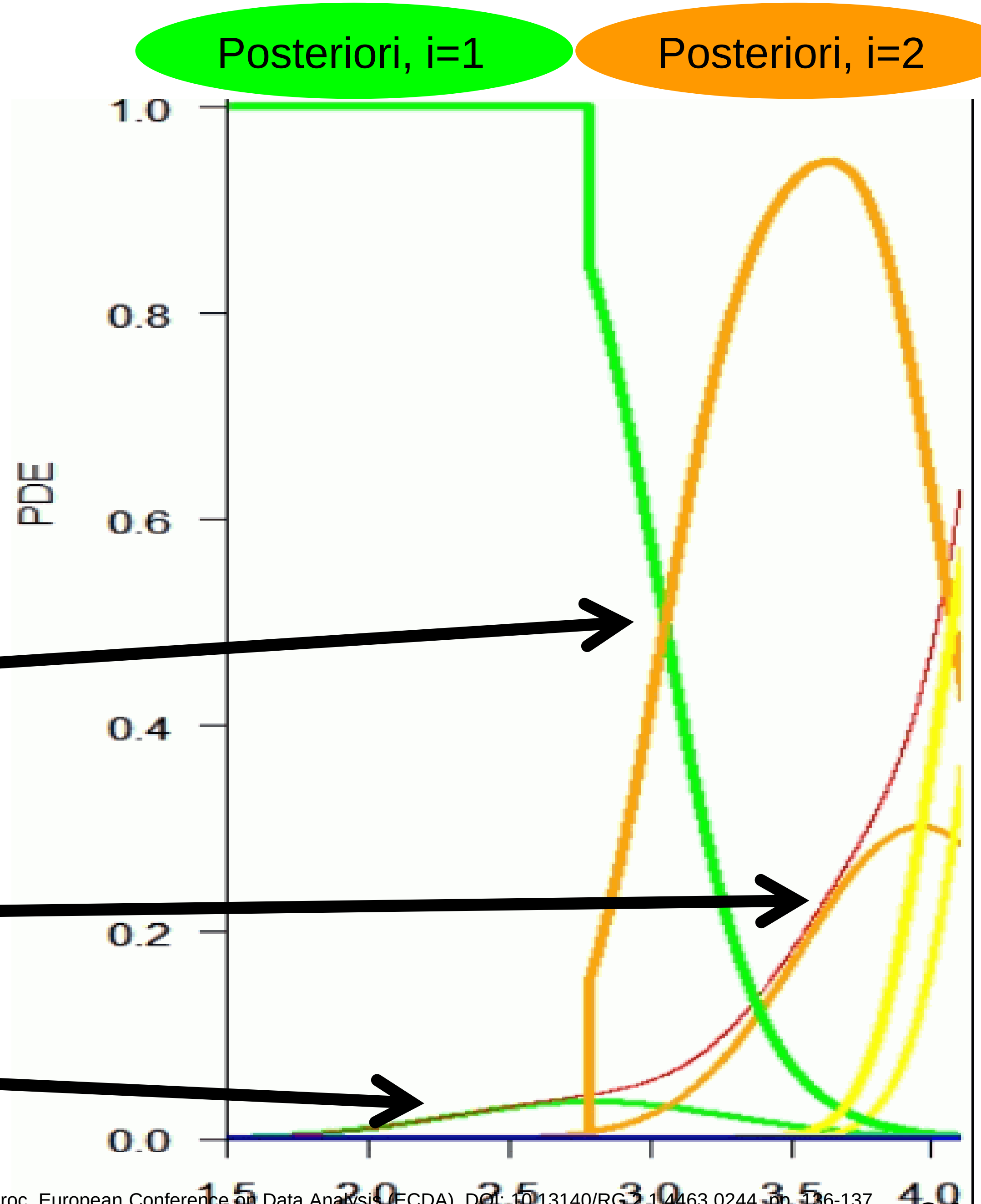
$$= \sum_{i=1}^{4} \mathrm{p}(c_i) * \mathrm{p}(\mathrm{x}|c_i)$$

(Details, see Bayes theorem)

Posteriori = 50%

Orange: Mixture Component i=2

Green: $N(m_1, SD_1)$ (i=1)



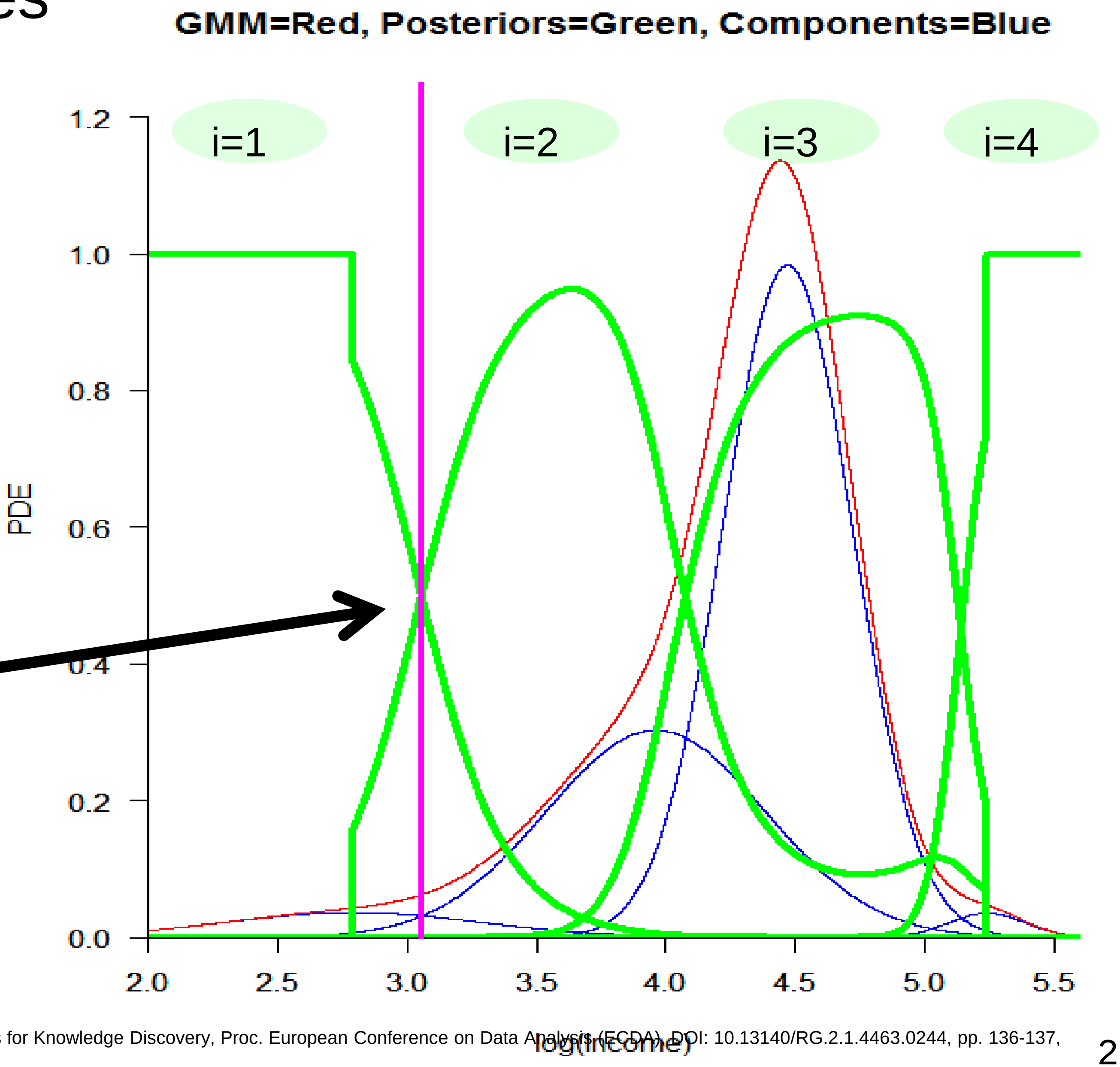
Thrun, M. C., & Ultsch, A. : Models of Income Distributions for Knowledge Discovery, Proc. European Conference on Data Analysis (ECDA), DOI: 10.13140/RG.2.1.4463.0244, pp. 136-137, Colchester, 2015.

Exact Boundaries

Green: Calculated posteriori of mixture components $i = 1, \ldots, 4$

Posteriori = 50%

$\Rightarrow$ Bayes Boundary between $i = 1$ and $i = 2$ (magenta)



Thrun, M. C., & Ultsch, A. : Models of Income Distributions for Knowledge Discovery, Proc. European Conference on Data Analysis (ECDA), DOI: 10.13140/RG.2.1.4463.0244, pp. 136-137, Colchester, 2015.