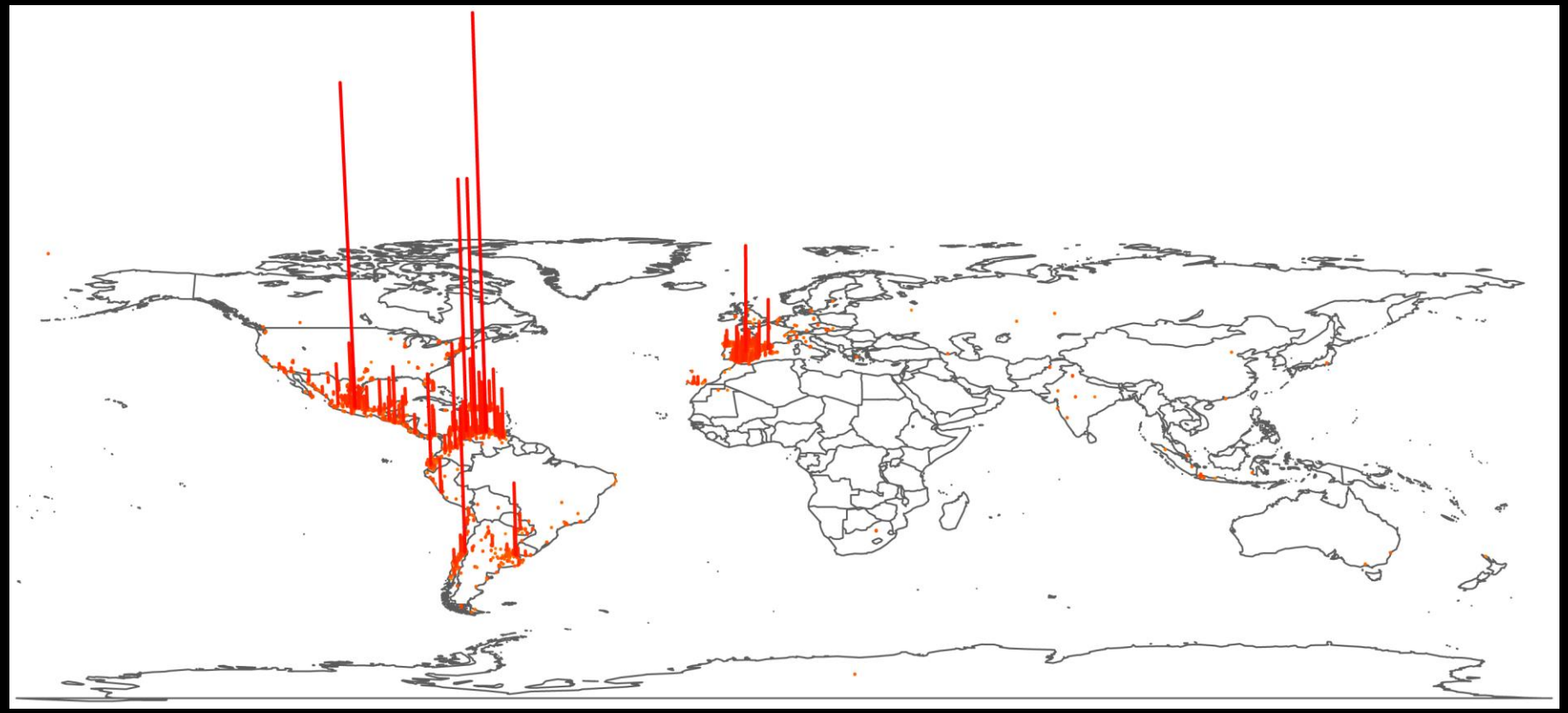
ATS1208
Digital Humanities:
Concepts, tools and debates

Week 10: Impacts - Scholarship



Overview

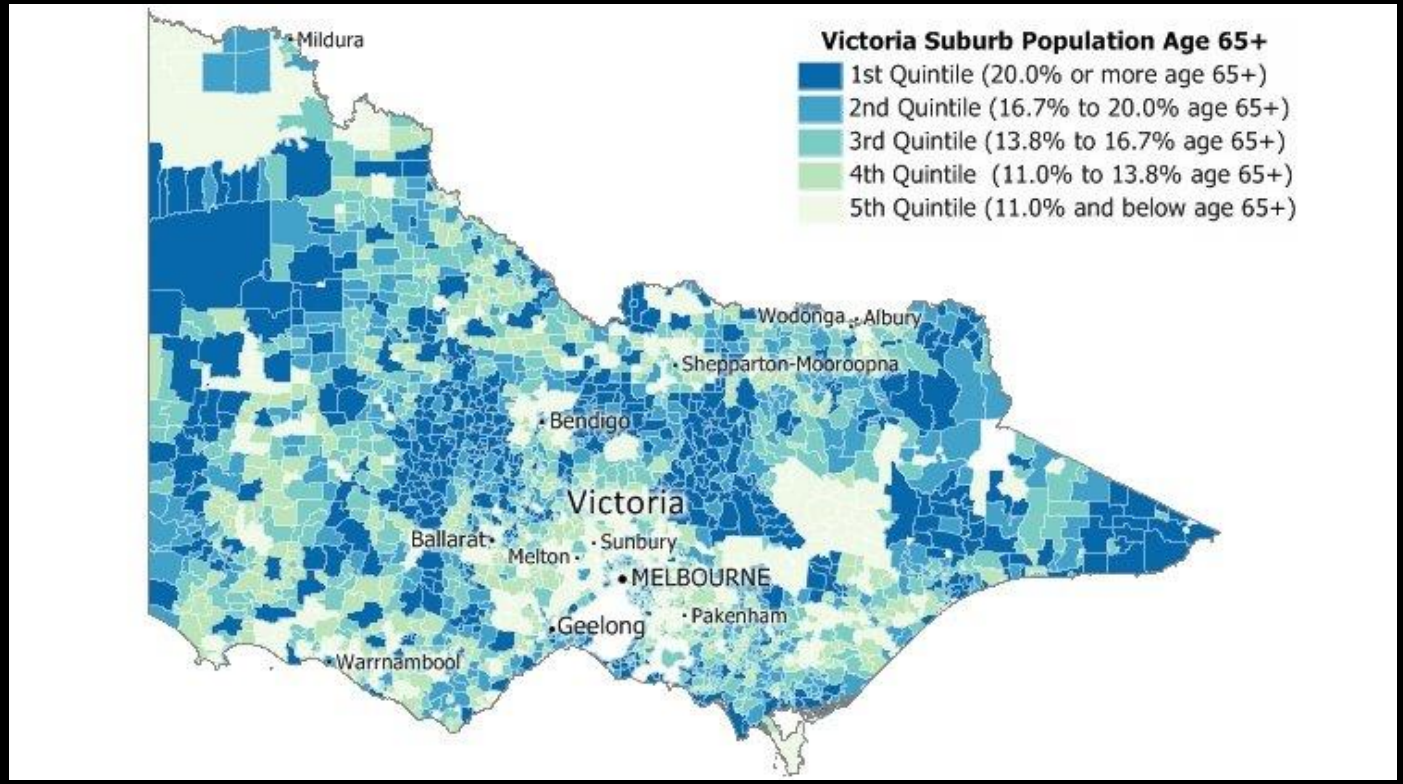
- Space and digital scholarship
  - GIS
- Space as a variable
  - Comparing two places: Thomas and Ayers
  - Comparing many places: Goncalves and Sanchez
- Space as an organising dimension: Belgum, Handley and Bott
- Workflows

Space and digital scholarship

- Space and place can be a greater part of scholarship because:
  - Much easier to incorporate images, maps etc.
  - GIS allows geographic analysis of a wide range of data

Geographic Information Systems (GIS)

- Systems which allow analysis of geocoded data
- For example:
  - Census data gives demographic details
  - Add geocoding at desired level of detail
  - Analyses can be carried out by geographic location
- Very easy to present results as a map or maps
  - But this is not essential for GIS analysis



GIS in the three articles

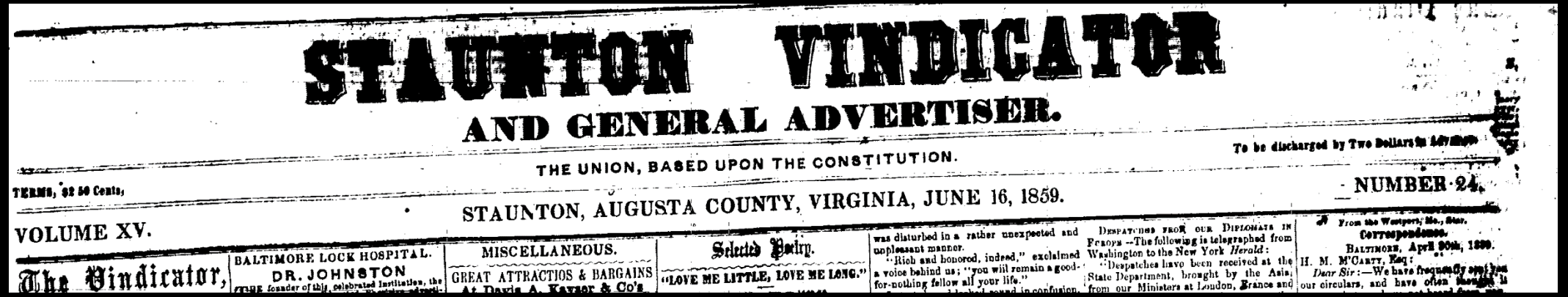
- Thomas and Ayers took large amounts of historical data and made it usable for GIS analysis
- Goncalves and Sanchez used data which came already geocoded
  - Carried out GIS-style analysis on it
- Belgum et al use geoparsing to make data available for geographic analysis
  - Their end product is a hand-tooled GIS system

Thomas and Ayers – what they did

- Look at two communities which are as similar as possible
- Except for presence/absence of slavery
- Chosen communities had to be:
  - Close together (both in Shenandoah Valley region)
  - Similar economic possibilities

Thomas and Ayers: data sources

- Article is based on a rich array of data:
  - Text sources – newspapers, diaries, letters etc.
  - Census records
  - Contemporary maps
  - (others)
- Data made digital and transformed as necessary for GIS analysis



STAUNTON VINDICATOR

AND GENERAL ADVERTISER.

THE UNION, BASED UPON THE CONSTITUTION.

To be discharged by Two Dollars in Advance.

TERMS, $2 50 Cents,

STAUNTON, AUGUSTA COUNTY, VIRGINIA, JUNE 16, 1859.

NUMBER 24.

VOLUME XV.

The Vindicator,

BALTIMORE LOCK HOSPITAL.

DR. JOHNSTON

MISCELLANEOUS.

GREAT ATTRACTIOS & BARGAINS

Selected Poetry.

"LOVE ME LITTLE, LOVE ME LONG."

was disturbed in a rather unexpected and unpleasant manner.

"Rich and honored, indeed," exclaimed a voice behind us; "you will remain a good-for-nothing fellow all your life."

DESPATCHES FROM OUR DIPLOMATS IN FRANCE.—The following is telegraphed from Washington to the New York *Herald*:

"Despatches have been received at the State Department, brought by the Asia, from our Ministers at London, France and

From the Westport, Mo., Star.

Correspondence.

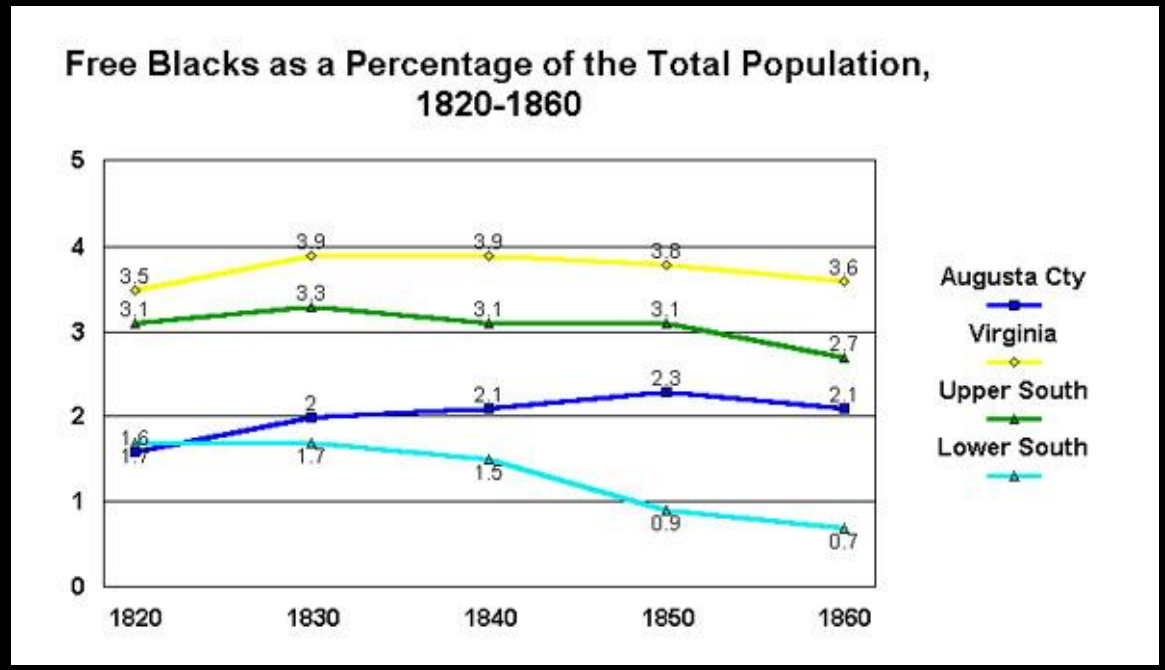
BALTIMORE, April 20th, 1859.

H. M. M'CARTY, Esq:

Dear Sir:—We have frequently sent you our circulars, and have often

Thomas and Ayers: access to data

- Selected data directly supports argument
  - Available through Points of Analysis tab
  - More data is available on associated website
- Underlying data is much more accessible than in conventional (printed) work
- A lot of the data is about place and space



Thomas and Ayers: presentation

- The authors set out to be innovative
- Features:
  - Making data easy to reach
  - Allowing reader to choose a path through material
  - No print version exists
    - The physical journal has an overview only (8 pages long)
    - No downloadable pdf or anything similar
- Radical approach, still impressive after 15 years

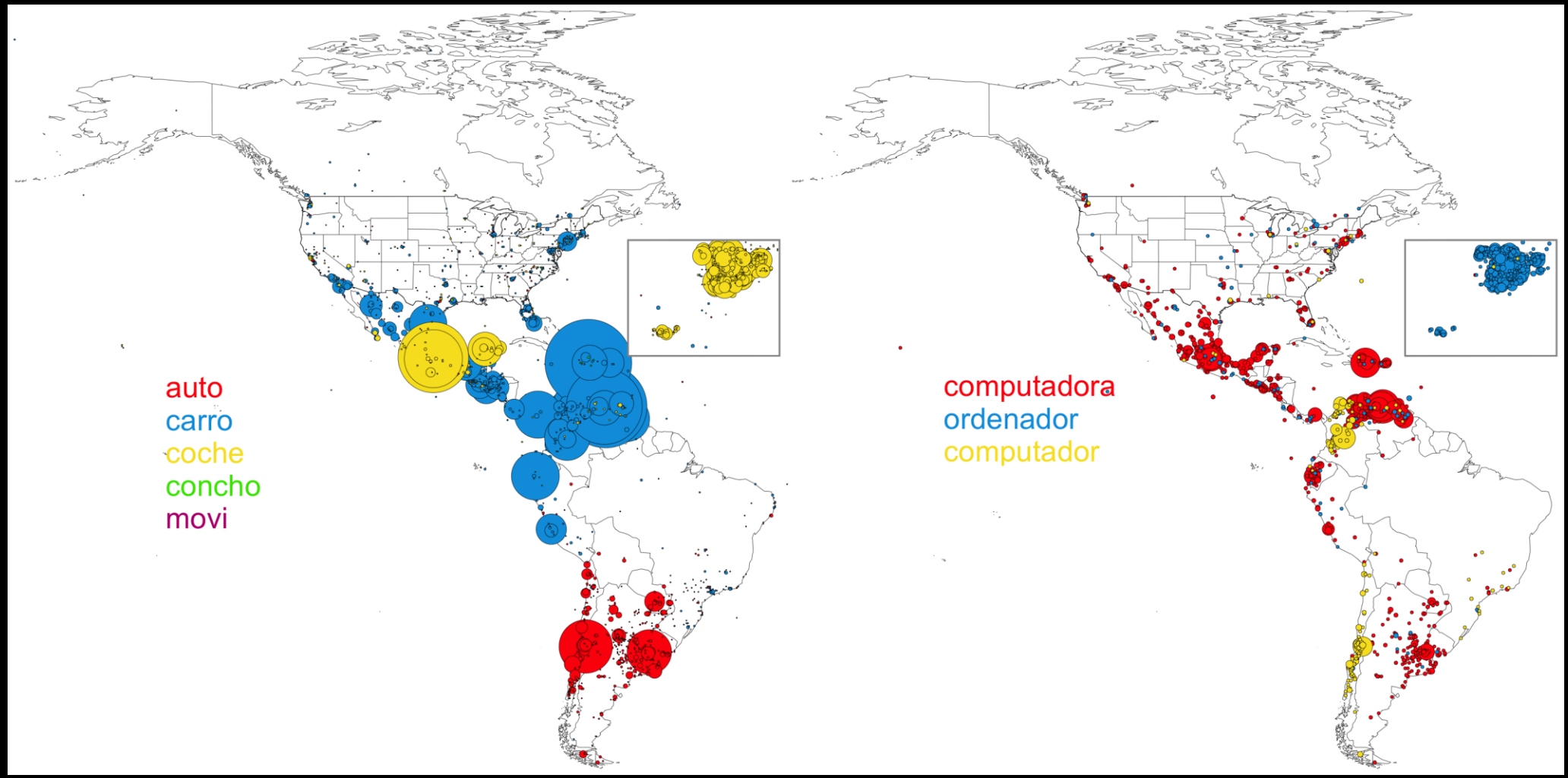
Goncalves and Sanchez: what they did

- Data is tweets in Spanish
- All geolocated tweets in Spanish collected for 2 years
- Total dataset – $> 5 \times 10^7$ tweets
- Built a list of concepts each represented by several possible words
- 7.5×10^5 tweets in the full dataset contained one of these words

Goncalves and Sanchez: the maths

- Geographic division into cells 0.25^{o} x 0.25^{o} (c $25km^2$ at the equator)
- Dominant word for each concept in each cell calculated by simple majority rule
- Results in matrix of 1135 rows (geographic cells) and 131 columns (concepts)
- Each point in matrix has value of 1 or 0
- Example of what it looks like on next slide

Cars and computers



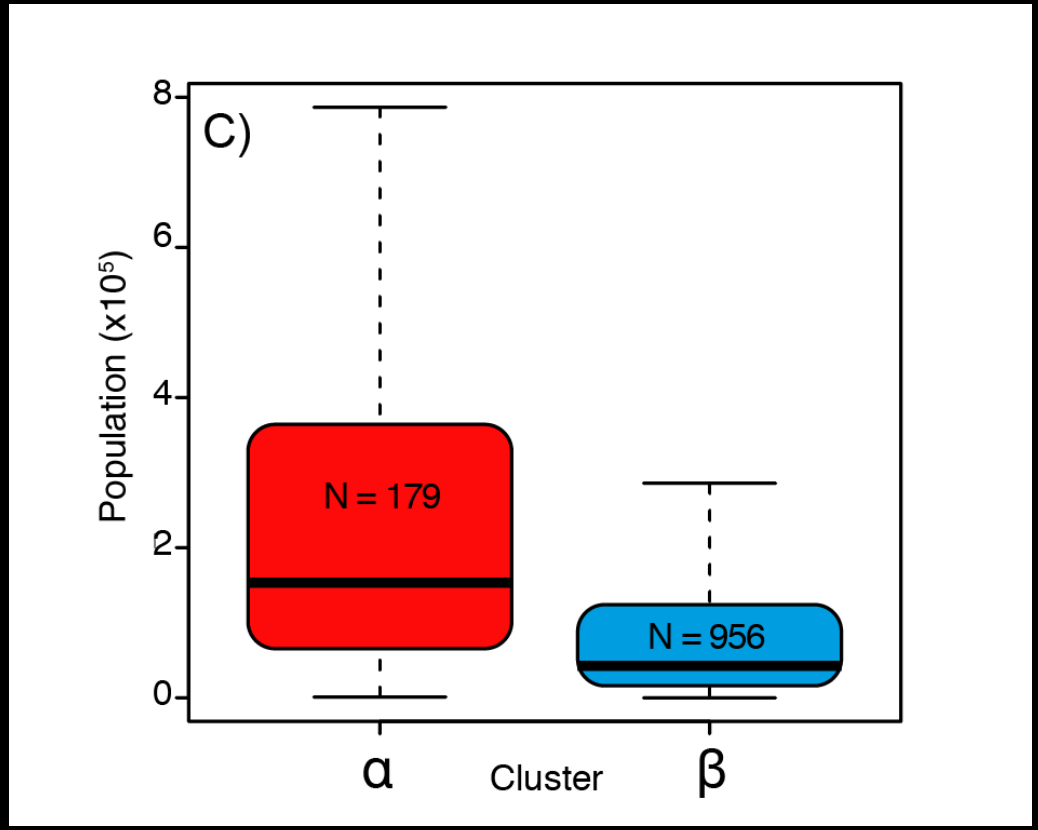
Machine learning

- Clustering of data was done with machine learning method
- Two steps:
  - Principal Components Analysis to reduce dimensionality (first 40 components included 94% of the variance)
  - Clustering with K-means algorithm
  - Best solution is 2 clusters

What it all means

- First look at geographic make up of clusters suggested population density was relevant
- Data on population density for each geographic cell imported from an existing dataset
  - LandScan 2007 High Resolution global Population Data Set
  - Using a GIS data source
- Next slide shows boxplot comparing population density for two clusters

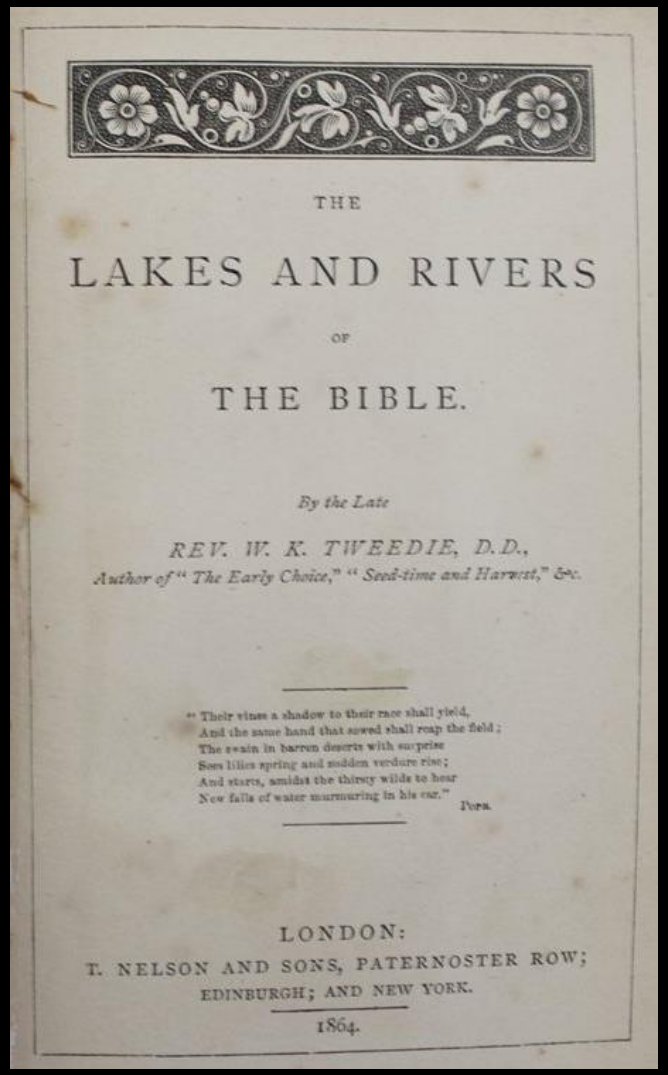
Superdialects and population



What it all means

- Standard story has been that major split in Spanish is between European variety and New World variety
- Goncalves and Sanchez show that major split is between an international urban variety and rural varieties
  - Cluster β can be split into several regional varieties
- Caveats:
  - Based on lexical variation – pronunciation variation is also important
  - Population may not be representative (younger, good knowledge of technology,….)

Belgum, Handley and Bott: what they did



THE

LAKES AND RIVERS

OF

THE BIBLE.

By the Late

REV. W. K. TWEEDIE, D.D.,

Author of "The Early Choice," "Seed-time and Harvest," &c.

"Their vines a shadow to their race shall yield,
And the same hand that sowed shall reap the field;
The swain in barren deserts with surprise
Sees lilies spring and sudden verdure rise;
And starts, amidst the thirsty wilds to hear
New falls of water murmuring in his ear."

Pope.

LONDON:

T. NELSON AND SONS, PATERNOSTER ROW;

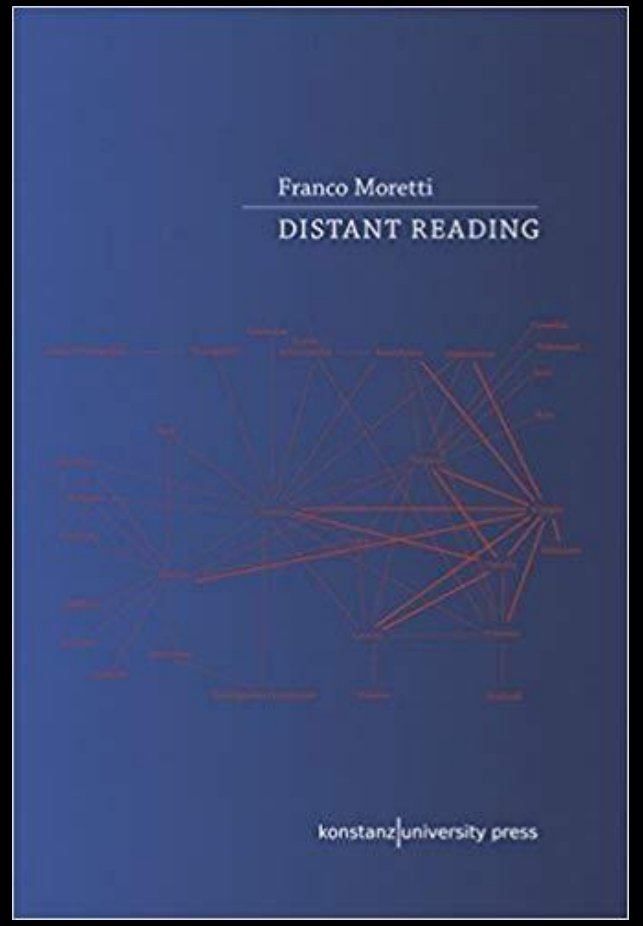
EDINBURGH; AND NEW YORK.

1864.

- Starting point is bibliographies of travel books published in C19
  - 3000 works published in Britain
  - Titles taken as basic data
- Each work was located geographically on the basis of places mentioned in the title
- Results can be accessed via interactive map

Distant reading

- Idea introduced by Franco Moretti
- Opposed to traditional literary analysis: close reading
- Moretti published an essay on the titles of English novels in C18 and C19
- He correlates changes and linguistic patterns with changes in society and publishing
- Current study is less ambitious – first question is which places were written about (and when)



First attempt

- Used bibliography of French travel writing
- Place names identified by hand (student assistants)
- Relied on coder knowledge and reference works
- Processing 100 entries took on average one hour
  - Plus time to review and correct errors

Geoparsing

- Takes unstructured references to locations
- Tries to resolve them to unambiguous identifiers (e.g. co-ordinates)
- Steps:
  - Identifying references in text (cf. Named Entity Recognition)
  - Match text reference to entry or entries in geographic database
  - Resolve ambiguities if possible
- Problems:
  - Variant names (archaic names and spellings)
  - Ambiguities – Paris, France v. Paris, Texas

Geographic database

- GeoNames is a very large digital gazetteer
- Stores variant names
- But no information on frequency of use
- This project used Wikipedia
  - Entries on places typically include geographic identifier (latitude and longitude)
  - Good coverage of higher level topography
    - Not many travel books have e.g. village name in title
  - Search history can be used for disambiguation
    - Paris, France is a common search term, Paris, Texas is not

Searching Wikipedia

- Procedure started with 5 word strings from title, then 4 word strings and so on down to individual words
  - Aim was to match e.g. New York City before New York
- Match of string to Wikipedia entry with geographic coordinates taken as identification of a place name
- Place name and co-ordinates added to bibliography database
- Manual checking as final step, with a map interface

Remaining problems

- Small number of cases where Wikipedia information did not align well with book
  - Manual input required
- Assigning point locations to regions is a general problem
  - Broad regions in titles (e.g. *The Islands of Greece)* were hard to identify

Results

- Online tool to visualise the data
- Provides details of individual works
- Possibility to view variation over time
- Extensible system – more data can be added

Workflows

- Each article details methods
- In each case, initial data needed to be manipulated in some way for further use
- Key element in digital humanities scholarship:
  - Finding structure in fully/partially unstructured data
  - Transforming data to make structure clear and available for analysis

What we spend our time on…..

References

- Belgum, Kirsten, Keith Handley & Rachel Bott. 2018. Mapping travel writing: a digital humanities project to visualise change in nineteenth-century published travel texts. *Studies in Travel Writing* 22(3). 306–324. doi:10.1080/13645145.2019.1575765.
- Gonçalves, Bruno & David Sánchez. 2014. Crowdsourcing Dialect Characterization through Twitter. (Ed.) Tobias Preis. *PLoS ONE* 9(11). e112074. doi:10.1371/journal.pone.0112074.
- Thomas III, William G. & Edward L. Ayers. 2003. The differences slavery made: A close analysis of two American communities. *The American Historical Review*. http://www2.vcdh.virginia.edu/AHR/.