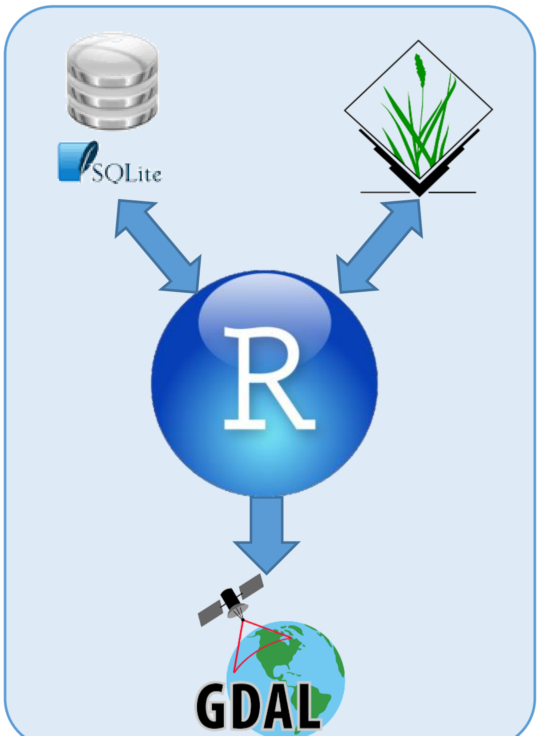
High Resolution, Multi-Year Compatible Dasymetric Models of US Population

Dasymetric modeling implementation in R



Anna Dmowska[1,2] (*dmowska@amu.edu.pl*)

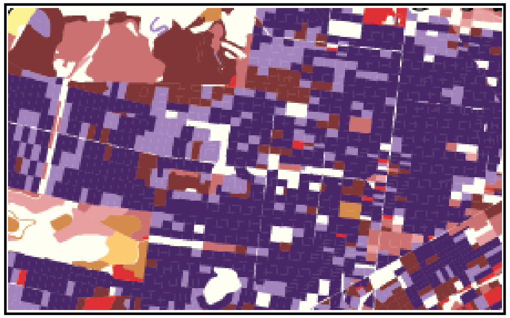
Tomasz Stepinski[1] (*stepintz@ucmail.uc.edu*)

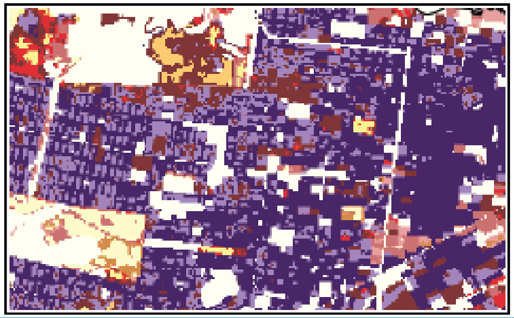
Paweł Netzel[1] (*netzelpl@ucmail.uc.edu*)

[1] Space Informatics Lab, University of Cincinnati, US
[2] Adam Mickiewicz University, Poznan, Poland

GIScience 2016, 27.09.2016-30.09.2016, Montreal, Canada

Demographic data

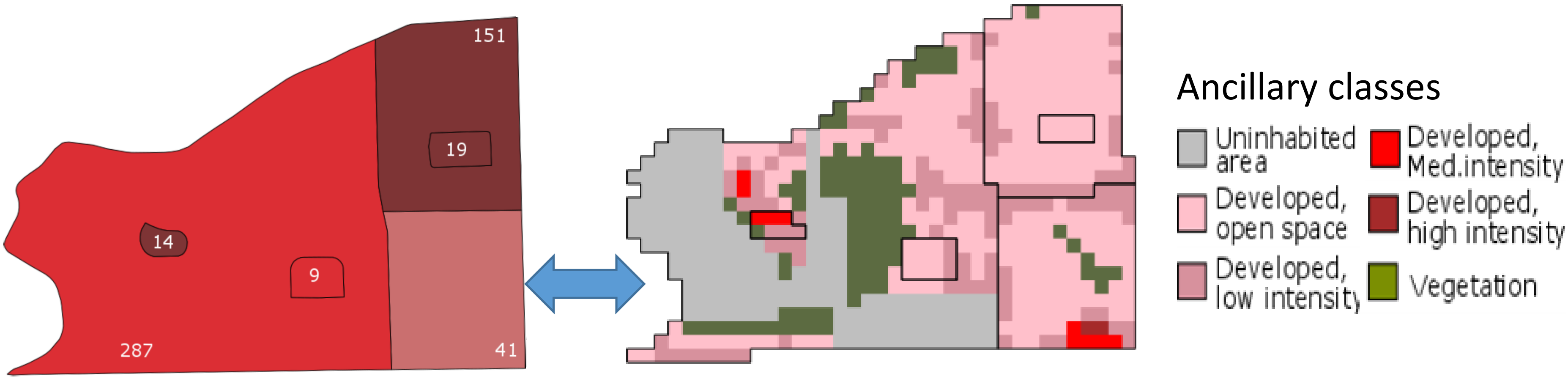




ELEMENT	DATA AGGREGATED TO UNITS	GRIDDED DATA
Availability	**Available** in different level of aggregation **for each** decenial **Census for the entire U.S.**	• Prepared primarily on a **local scale** • **Not available as multi-year compatible datasets**
GIS format	**vector** (shapefile) + **attribute table**; **difficult to work** with large shapefile files	**raster**; **easy to work** with large raster files
Spatial resolution	**dependent on the choice of Census units** and spatially varying; low in rural areas	**high and spatially constant**; defined by the size of the cell
Uniformity	mapped population **distributed uniformly** within each Census unit	mapped population density **changes continuously from cell to cell**
Temporal change	the extents of **Census units change with time**, which makes difficult year-to-year comparison	grid **enables direct** cell-to-cell **temporal comparison**

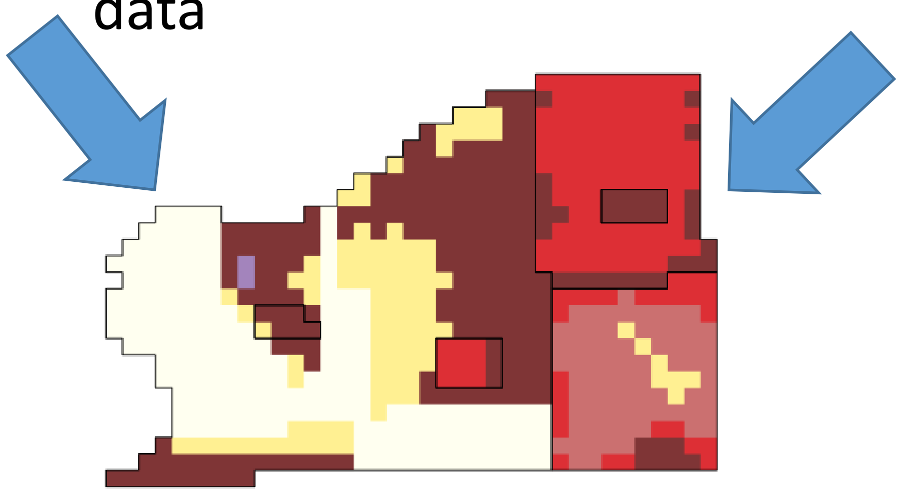
There is **a need to develop hi-res multi-year compatible** population datasets, which can be used to perform analysis **for large areas.**

Dasymetric modeling



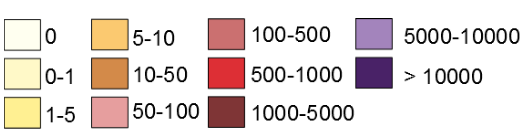
Aggregated demographic data

Ancillary data



Dasymetric model

People/km²



Dasymetric modeling refers to a process of **disaggregating spatial data to a finer unit** of analysis, **using additional** (or **ancillary**) **data** to help refine locations of population or other phenomena (Mennis 2003).

What makes it difficult to perform dasymetric modeling for large areas?

- **Size of dataset**
  - Size of input data: 11 milions of aggregated units (U.S. Census 2010)
  - Size of output data: 8 bilions of cells (30x30m grid)
- The **limited availability of high resolution ancillary data** for large areas
  - Ancillary data must be available for the entire area in **uniform fashion and quality**
  - Ancillary data must be **comparable between years**
- The need to develop an **efficient, fully automated algorithm** to work with large datasets, which will allow to perform calculations within a reasonable time

Our project

The goal of this project is to provide an **open and convenient access** to high resolution, **multi-year compatible** population data.

1. **Developing software** to perfom dasymetric modeling for large datasets

2. **Preparing hi-res**, multi-year compatible population g**rids**:
Spatial resolution: 30 m, Time: 1990, 2000, 2010

3. Developing **GeoWeb application** to provide **open access** to population grids

Developing software – dealing with large datasets

GIS SOFTWARE	SOFTWARE BUILT FROM SCRATCH
+ contains **a lot of „ready to use”** tools designed to solve a specific task	+ computationally **efficient** for large datasets
- it is **difficult to extend** if the problem goes beyond its built-in capabilities	+ we decide about its functionality and flexibility
- computationally **inefficient** for large datasets	- requires **advanced programming** skills (i.e C, C++)

WORKING IN R ENVIRONMENT

+ Allow to build **efficient**, **flexible** and **fully automated** computational environment to work with large dataset **without advanced programming** skills.

+ R is a comprehensive computational environment that includes **libraries to work with different types of data**: ***geospatial data*** (sp, rgrass7, raster, rgdal), ***standard relational databases*** (DBI, RSQLite).

+ **Main advantages of using R over GIS software** are: less processing steps are required, no intermediate layers, increased flexibility and automation.

How our algorithm works?

STEP 1a
Preprocessing Census data

STEP 1b
Preprocessing ancillary data

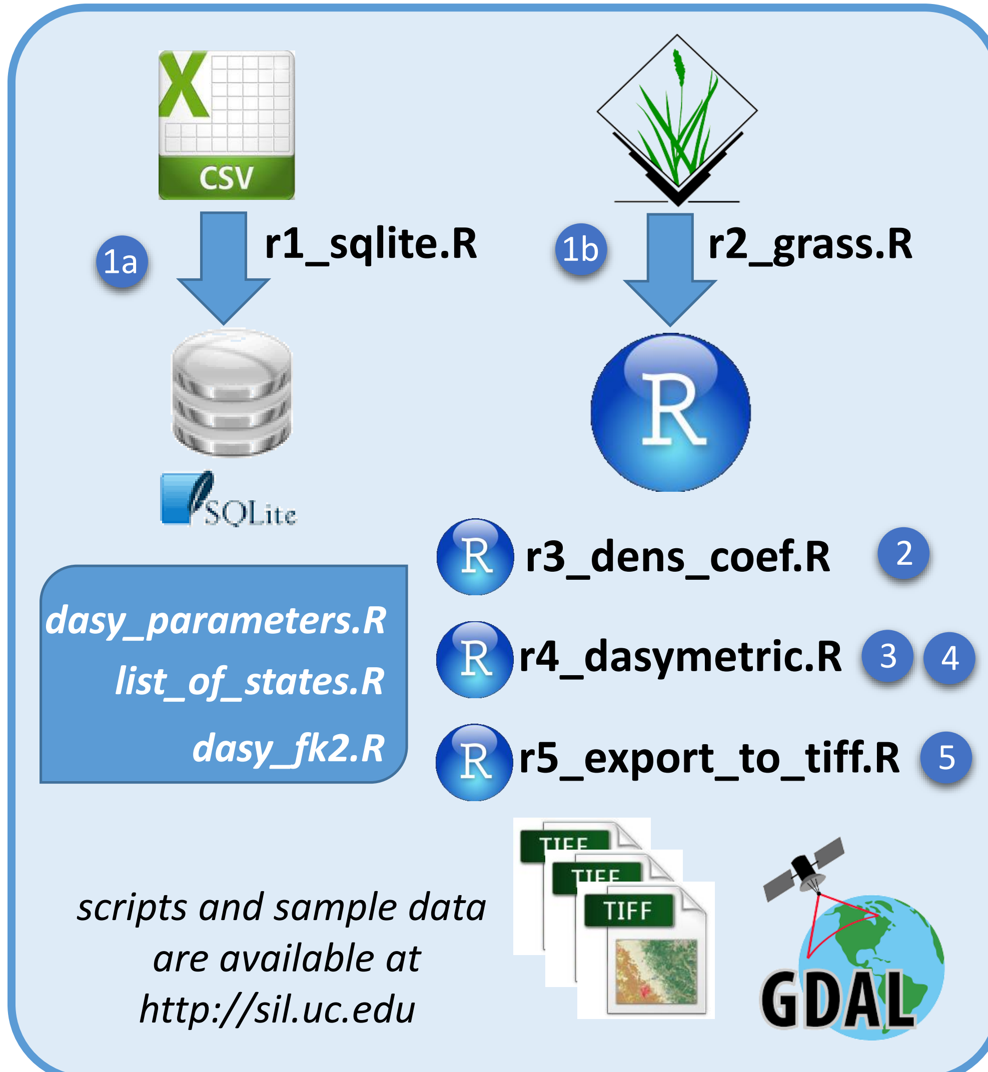
STEP 2
Determine relationship between demographic and ancillary data

STEP 3
Perform dasymetric modeling

STEP 4
Propagate dasymetric model for geospatial grid

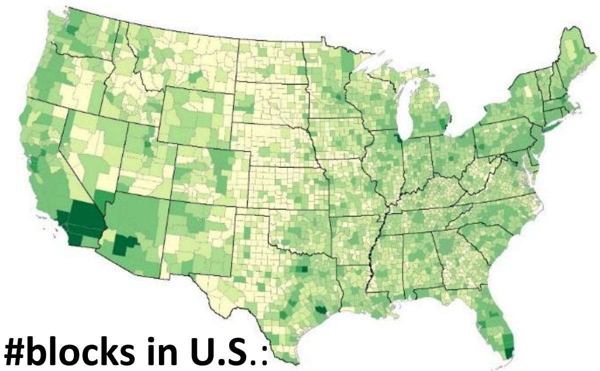
STEP 5
Postprocessing: Prepare hi-res population maps for the entire U.S.

Scripts for dasymetric modeling calculations



Applying dasymetric modeling to U.S.-wide data

DEMOGRAPHIC DATA



#blocks in U.S.:
~7,15 milions (1990), ~8,2 milions (2000), ~11,15 milions (2010)

The **1990, 2000, and 2010 decennial Censuses** data aggregated to **block level**.

This data consist of **two components**:

- shapefiles (TIGER/Line Files) indicating **blocks' geographical boundaries** – available for **each state** separately
- summary text files which list **population** data **for each block** – available **for the entire U.S**. as one tabular file

ANCILLARY DATA

Data size:
104424 x 161190 = 16 832 104 560 cells
(no-null cells: 8 651 173 750)

Land cover data - the **only high resolution ancillary data** available **for the entire U.S.** in uniform fashion and quality.

NLCD products:

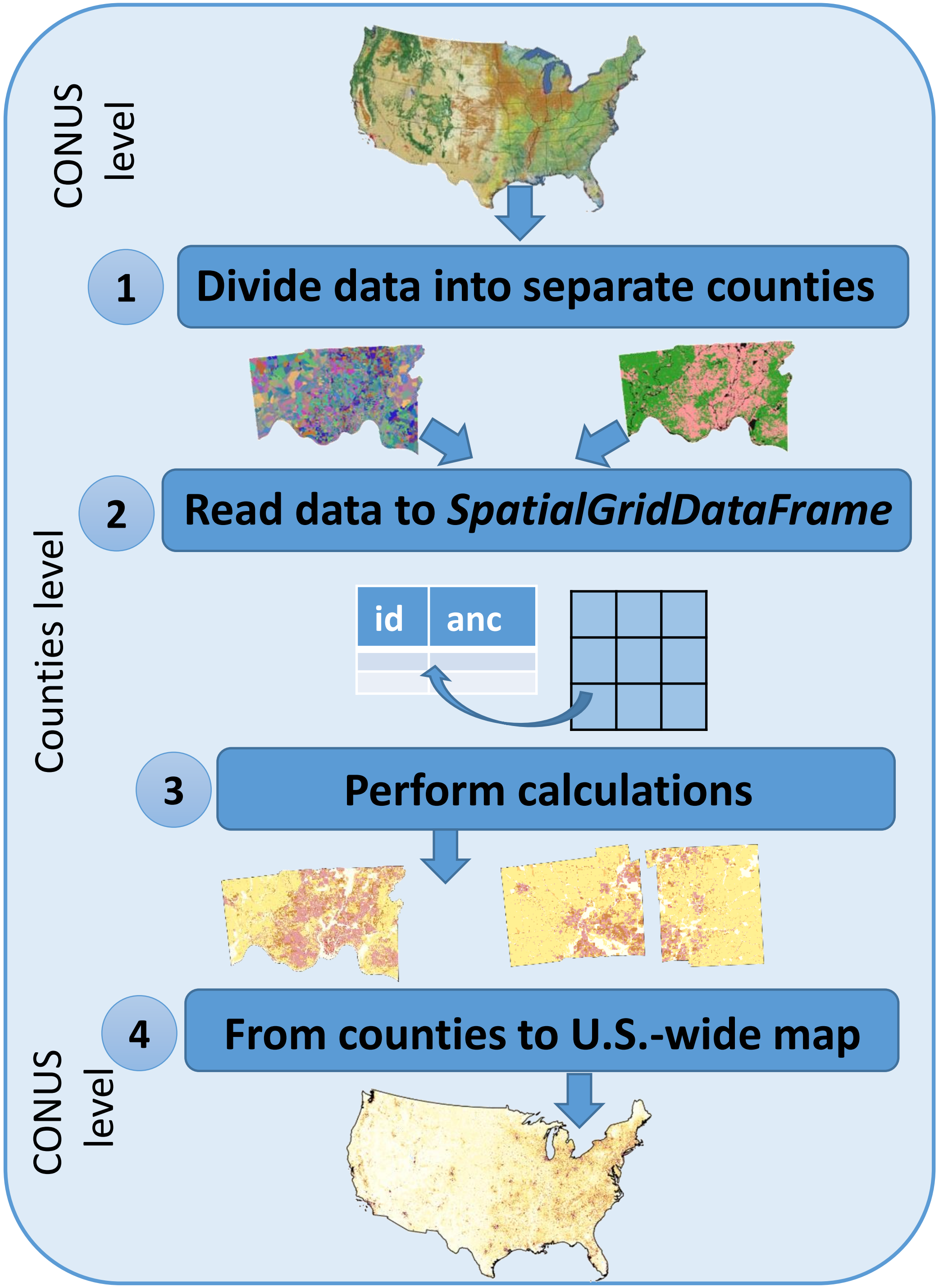
- NLCD 2001
- NLCD 2011
- NLCD 1992/2001 Retrofit Land Cover Change Product

NLCDs are **reclassified to 3 classes** (uninhabited, urban, vegetation) to preserve comparability between years.
NLCD1992 has incompatible legend with next editions and can't be used for change analysis. Instead NLCD 1992/2001 Retrofit Land Cover Change Product was developed.

Handling large dataset in R

1. To manage data storage requirements and to better control the time of computation we **divide U.S. into separate counties**.
    - We used region concept in GRASS GIS for computationally efficient division of U.S. into separate counties.
2. Raster **data for each county is read into *SpatialGridDataFrame* object in R**
    - This structure allow to integrate information about its spatial content with Census data into a single relational model.
3. We **process each county separately**.
4. In the last step **maps for individual counties are joined into U.S.-wide map**

CONUS=conterminous U.S.



Our methodology

STEP 1a
Preprocessing Census data

STEP 1b
Preprocessing ancillary data

STEP 2
Determine relationship between demographic and ancillary data

STEP 3
Perform dasymetric modeling

STEP 4
Propagate dasymetric model for geospatial grid

STEP 5
Postprocessing: Prepare hi-res population maps for the entire U.S.

- **The population** in each block **is redistributed** to its cells using **block-specific weights** assigned to the cells having different ancillary classes.
- The weights are assigned based on two factors:
  - **relative density** of population for each ancillary class,
  - the **area of each block occupied by each class** (Mennis 2003).
- **Relative densities** are the representative densities normalized by the sum of representative densities for each class.
- **Representative population density** for each class is established using **a set of blocks** (selected from the entire U.S.) **having relatively homogenous land cover** (90% for developed classes and 95% for vegetation classes).
- **Population in each block's cell** = **number of people in the block multiply by** a **weight** assigned to the cell based on its ancillary class.

Preprocessing census and ancillary data

STEP 1a
Preprocessing Census data

STEP 1b
Preprocessing ancillary data

STEP 2
Determine relationship between demographic and ancillary data

STEP 3
Perform dasymetric modeling

STEP 4
Propagate dasymetric model for geospatial grid

STEP 5
Postprocessing: Prepare hi-res population maps for the entire U.S.

Census data

Population data for each block in the U.S.

Import to database

SQLite

Geospatial data

Blocks boundaries for each state

Ancillary data (NLCD for the entire U.S.)

Import to GRASS GIS

Preprocessing steps:

- Rasterize block's boundary shapefiles
- Reclassify NLCD into 3 classes
- Divide data into separate counties
- Export data to R (SpatialGridDataFrame)

Performed dasymetric modeling in raster data = significant gain in computational performance

Determine relationship between demographic and ancillary data

STEP 1a
Preprocessing Census data

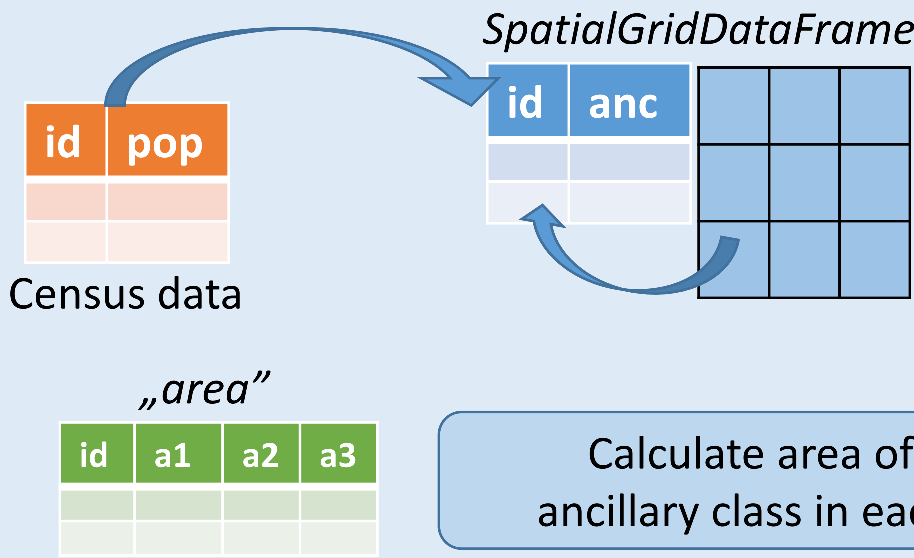
STEP 1b
Preprocessing ancillary data

STEP 2
Determine relationship between demographic and ancillary data

STEP 3
Perform dasymetric modeling

STEP 4
Propagate dasymetric model for geospatial grid

STEP 5
Postprocessing: Prepare hi-res population maps for the entire U.S.



Calculate area of each ancillary class in each block

Select block with relatively homogenous land cover class

Calulate representative population densities

Calulate relative density of population for each ancillary class

The **SpatialGridDataFrame integrate tabular** and **geospatial** data into a **single** relational model. In practice this means that all **the calculations are performed at the data frame** (tabular) level.

Performing dasymetric modeling in R

STEP 1a
Preprocessing Census data

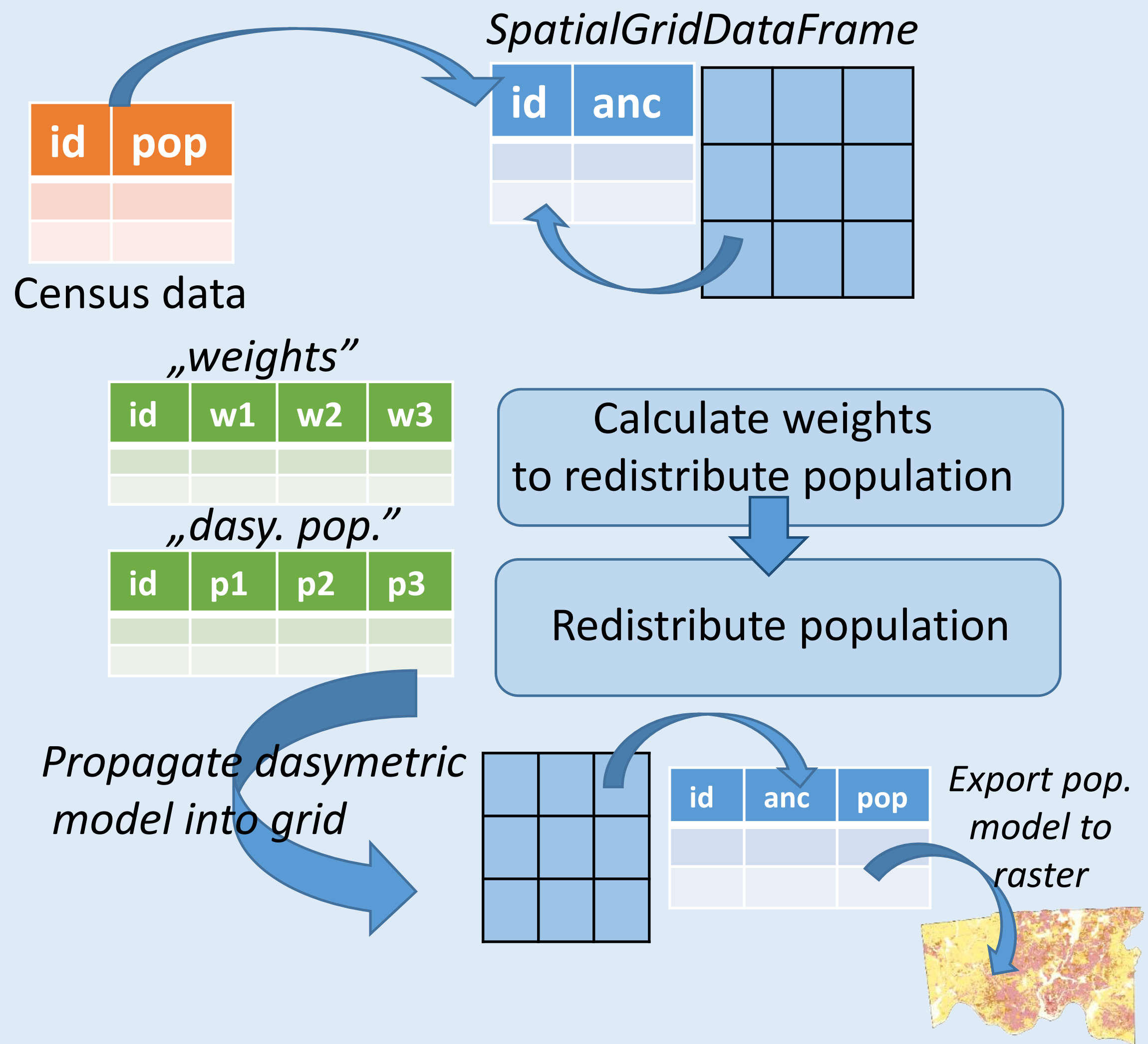
STEP 1b
Preprocessing ancillary data

STEP 2
Determine relationship between demographic and ancillary data

STEP 3
Perform dasymetric modeling

STEP 4
Propagate dasymetric model for geospatial grid

STEP 5
Postprocessing: Prepare hi-res population maps for the entire U.S.



The **SpatialGridDataFrame** allow to **integrate tabular** and **geospatial** data into a **single** relational model. In practice this means that all **the calculations are performed at the data frame** (tabular) level and next **propagate into a grid**.

Postprocessing: *From counties to U.S.-wide map*

STEP 1a
Preprocessing Census data

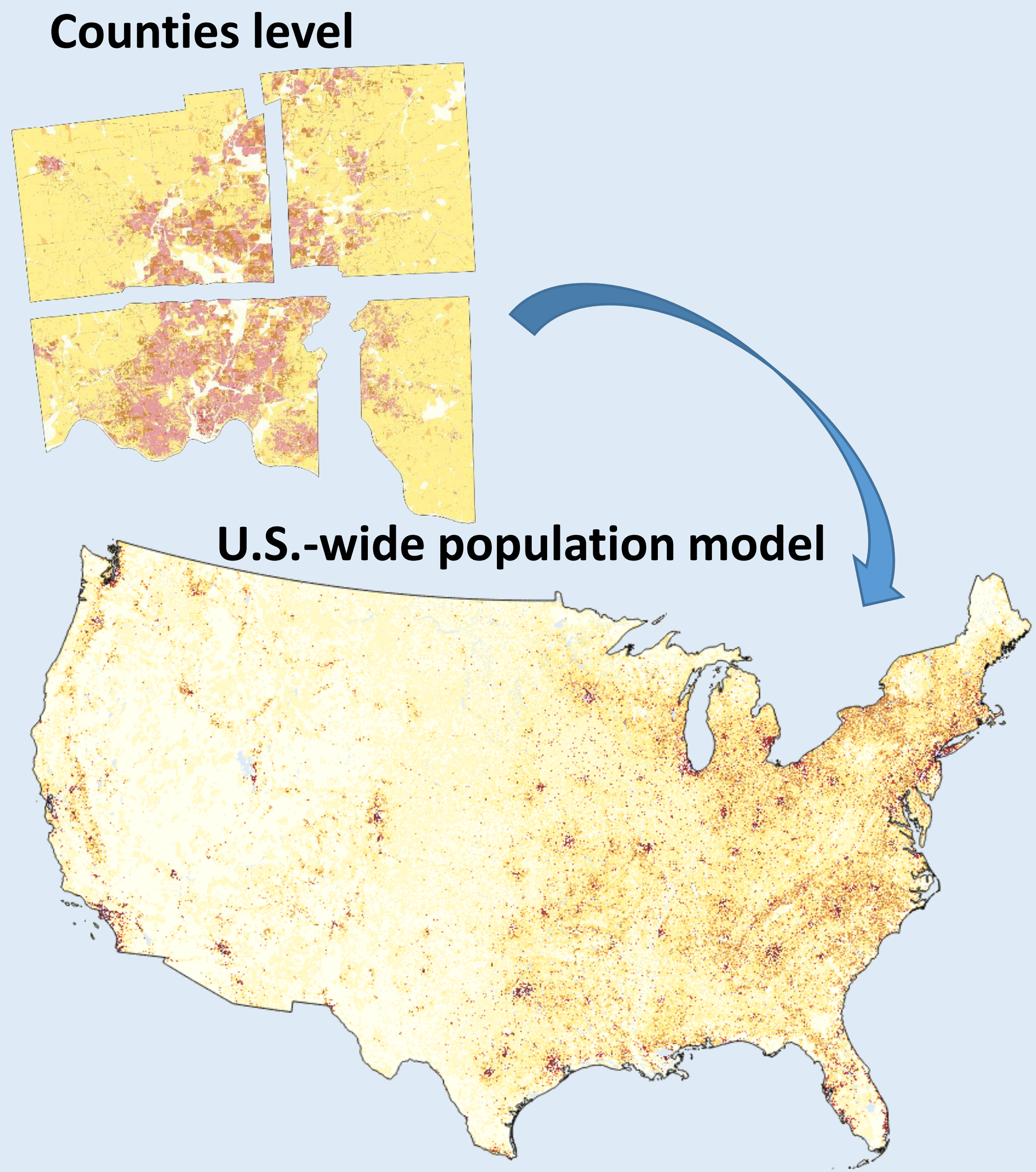
STEP 1b
Preprocessing ancillary data

STEP 2
Determine relationship between demographic and ancillary data

STEP 3
Perform dasymetric modeling

STEP 4
Propagate dasymetric model for geospatial grid

STEP 5
Postprocessing: Prepare hi-res population maps for the entire U.S.



In the last step maps for individual counties are joined into U.S.-wide map using GDAL tools.

Performance

GIS SOFTWARE VS R

GIS SOFTWARE (USGS ArcGIS toolbox)	R SCRIPTS
Hamilton County = 10808 Census blocks	
600 seconds (10 min.)	**14 seconds**

U.S. CALCULATION

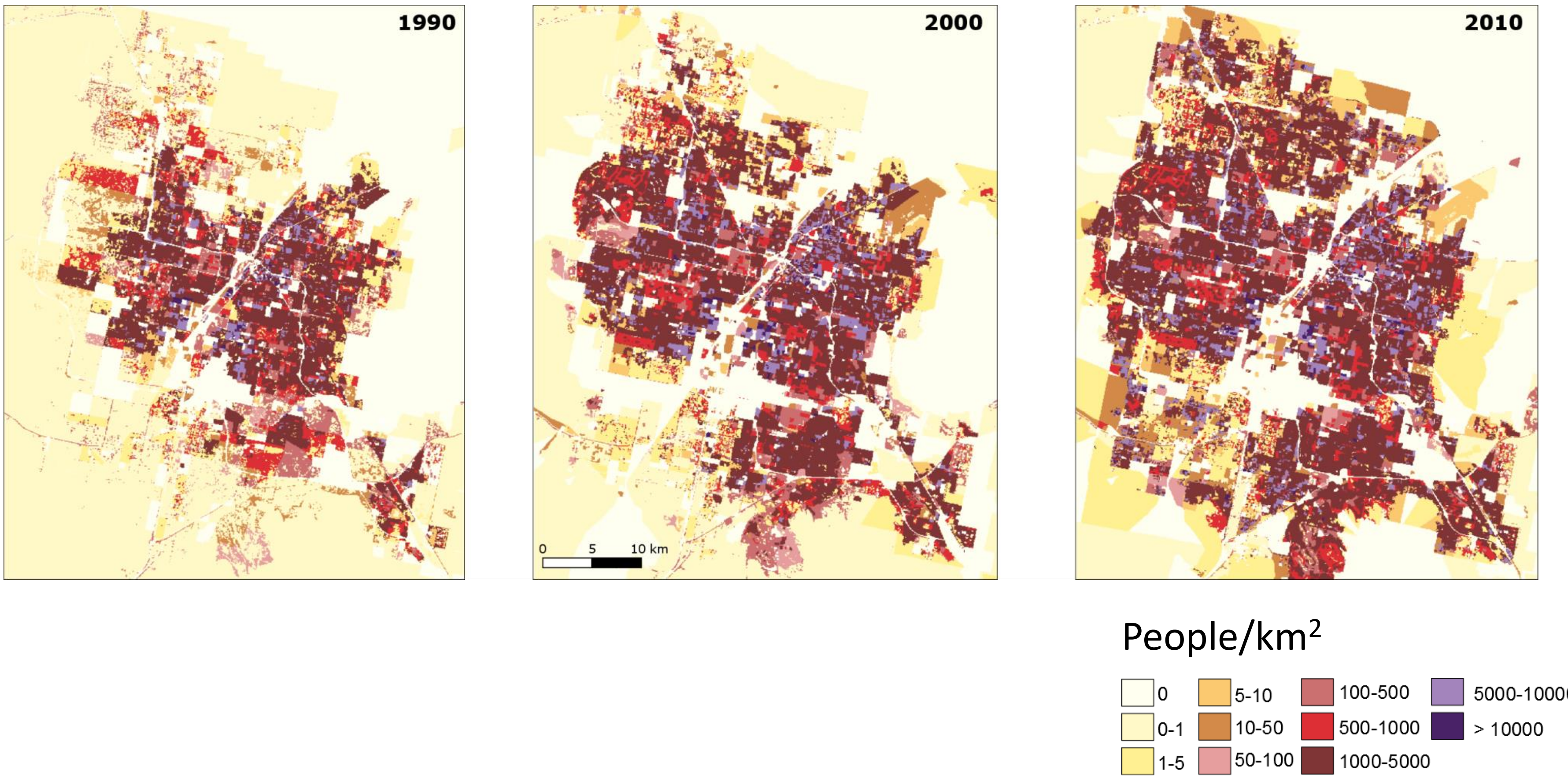
DATA	SIZE OF FILES
Nb. of blocks in U.S.	~7,15 milions (1990), ~8,2 milions (2000), ~11,15 milions (2010)
Size of output map in cells	16 832 104 560 (no-null: 8 651 173 750)

Processing steps	Calculation time
Preprocessing data	37 h
Determine relation between population and ancillary data, Perform dasymetric model	6 h
Create one map from counties	12 h
Overall time	**55 h**

All calculation was done using a PC computer with Intel 3.4GHz, 4-cores processor and 16 GB of memory running the Linux system.

High resolution multi-year compatible U.S. population model

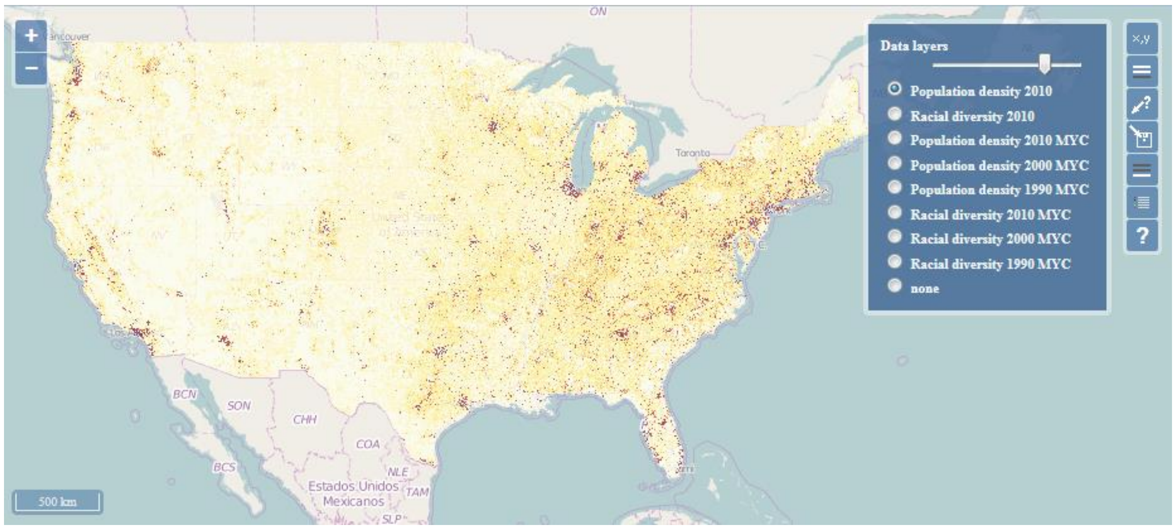
Population dynamics change for Las Vegas, NV based on hi-res grids for 1990, 2000 and 2010.



Providing open access to hi-res grids for the entire U.S.

SocScape – GeoWeb application

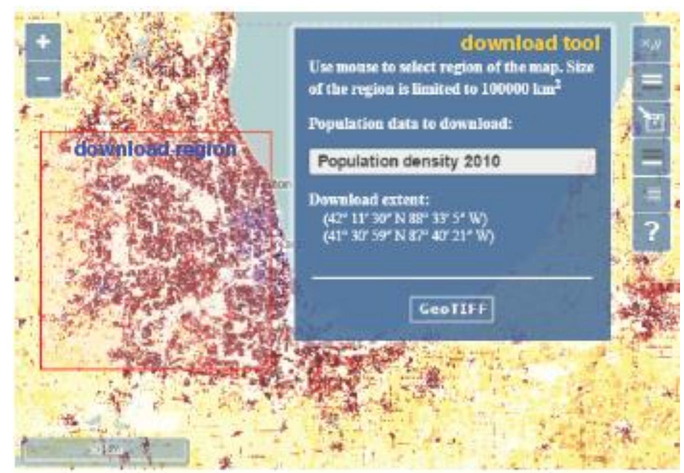
http://sil.uc.edu/webapps/socscape_usa/



Avalailable data:
Population distribution, racial diversity
Time: 1990/2000/2010

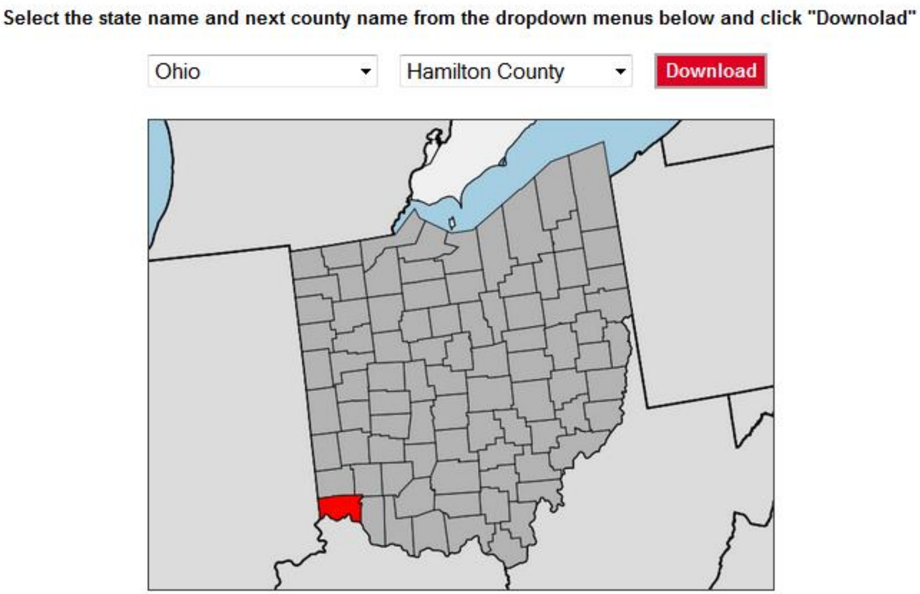
SocScape can be useful for:

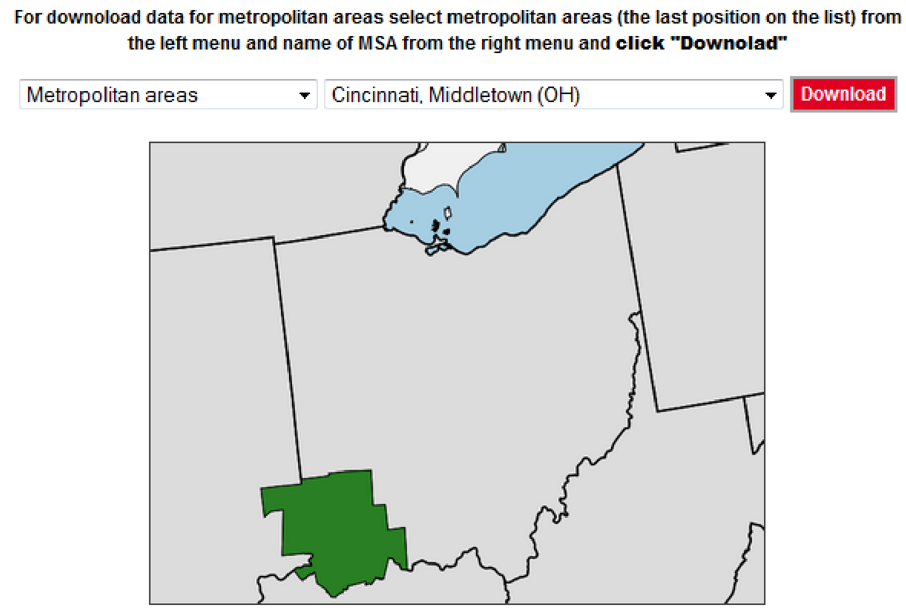
- Fast and **intuitive exploration of population density** and racial diversity **in different scales** (from the entire U.S. down to the street level)
- **Detecting spatial dynamics** of population density and racial diversity.
- **Downloading data** in Geotiff format for selected area (up to 100 000 km^2)



Providing open access to hi-res grids for counties and metropolitan areas.

http://sil.uc.edu/cms/index.php?id=socscape-data





Spatial extent:

- 3100 counties
- 363 metropolitan areas

Maps:

- Population distribution,
- Race distribution,
- Racial diversity

Time: 1990/2000/2010

Data for each unit (**county or MSA**) is organized as a **zip archive** containing data **for 1990, 2000 and 2010**.
Grids are saved as GeoTiff.

Conclusion

Our project to provide **open and convenient access to hi-res multi-year grids** of US population is now **completed**.

- We developed **30 m** resolution **grids of the U.S. population** in 1990, 2000, and 2010 using **a multi-year compatible dasymetric model**.
  - These grids are designed to assess population change across the conterminous U.S. at street-level spatial resolution.
- The model and its novel, **computationally efficient implementation in R** are presented.
- The **grids are available online** for interactive exploration and data download using especially developed GeoWeb application SocScape:

  http://sil.uc.edu/webapps/socscape_usa/

Acknowledgments

The authors wish to thank J.Jasiewicz for helpful comments on the development of computationally efficient implementation of dasymetric modeling algorithm in R.