Exploration of the Local Distribution of Major Ethnic Groups in the USA

Sebastian Kay Belle* University of Konstanz
Daniela Oelke† University of Konstanz
Sonja Oettl‡ University of Konstanz
Mike Sips§ Stanford University

ABSTRACT

Knowledge about the local distribution of major ethnic groups in the USA is an important source of information upon which the success of political and economic decisions may depend. Enhancing this information with additional attributes, such as the specific income or spoken languages, reveals interesting aspects on social constellations as well as their interdependencies. The presented visualization facilitates the intuitive exploration of such multidimensional data sets with references to geographical units. Thus, a quick insight into the inherent patterns and characteristic features is provided. Thanks to the high scalability of our visualization technique it can even be used with an iPod-resolution.

1 INTRODUCTION

Once every ten years the U.S. Census Bureau collects data of the people of the United States in a broad demographic survey. Part of the data, the Public Use Microdata Sample (PUMS 1%), can be used by everyone whenever a reliable basis for political, economic, or social decisions is required. Among others, the dataset contains information about nationalities, ethnicities, and the linguistic and cultural background of the investigated people. The information refers to small geographical units to allow to examine the spatial distribution of the values across the United States. Values for counties, states or even the whole USA can be obtained by aggregating corresponding subunits of the inherent hierarchy.

In our project we analyze the spatial distribution of eight major ethnic groups of the United States on different hierarchical levels. Moreover, we enhance the visualization with information about the median income of each group in the corresponding geographical unit and the regionally spoken languages to provide further indicators for social constellations and correlations.

2 THE CONCEPT

At first glance it may seem indispensible to use a map for the representation of data values with such a strong geographic reference. However, a major disadvantages of a map is that it is difficult to display the values of multiple ethnic groups or the spoken languages simultaneously. This becomes even more decisive if a comparison between different geographical units (such as different states or counties) should be possible.

Therefore, in our visualization we chose a matrix-like structure instead of displaying the values on a map. Each geographical unit is assigned to a column and the corresponding attribute values, such as the number of people belonging to the different ethnic groups, their income, or the spoken languages, are arranged in rows. Within each cell the single values are encoded with colors.

With this approach we may lose the intuitive perception of the geographic context that a map can provide. However, the regular grid structure also has some crucial benefits. First of all, we can easily display many dimensions at once. Second, different values can be compared efficiently. Since the distribution over the whole matrix can be perceived at a glance, patterns stick out immediately. Rearrangement of columns and rows according to the specific task or detected similarities may even increase this effect.

To account for the hierarchical structure of the data we provide a drill-down facility to view each state on a more detailed level.

*e-mail: belle@inf.uni-konstanz.de
†e-mail: oelke@inf.uni-konstanz.de
‡e-mail: oettl@inf.uni-konstanz.de
§e-mail: msips@graphics.stanford.edu

3 MAPPING OF THE VALUES TO COLOR

There are different methods to normalize the values of the matrix to map them to a color gradient (e.g., linear, square root, or logarithmic normalization). Furthermore, we have to decide if the normalization should be done on a row-by-row basis, a column-by-column basis, or with a whole matrix as one unit of normalization. Depending on the selected strategy, different questions can be answered. For example, the row-by-row normalization allows us to view quickly where each ethnic group tends to concentrate, whereas the column-by-column normalization results in visualizations that show the proportion of the different ethnic groups. Taking the whole matrix as a basis unit for the normalization has the advantage that general trends like patterns on state level can easily be recognized.

4 DATA ANALYSIS

Fig. 1 shows the values for number of households (brown), median income (violett) and household languages (green) on state level (left and right) and drilled down for the state New York (middle)[1]. The subfigures 1a to 1c are calculated with different normalization strategies.

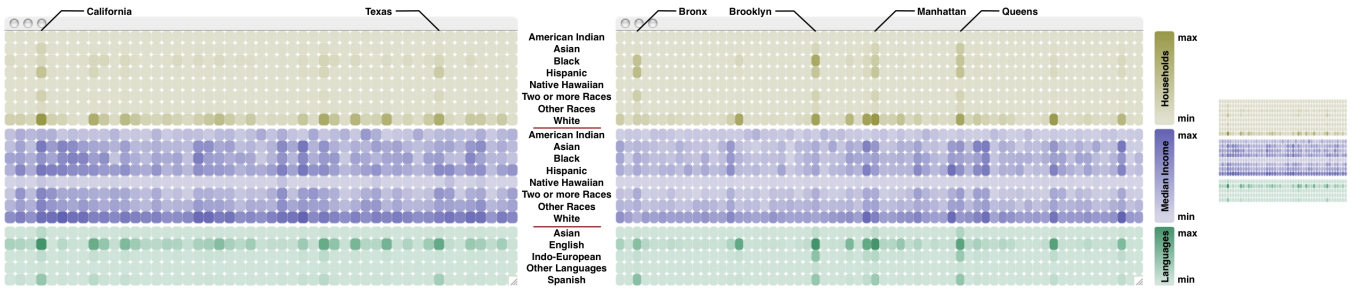
In fig. 1a (left) patterns on state level can easily be perceived. For example, it is obvious that California and Texas are states with a high total population.

With a row-by-row normalization as in fig. 1b the distribution of each ethnic group across the USA can be analyzed. Unsurprisingly, the states with a high population also tend to be the ones in which most of the people of each ethnic group live. However, it is interesting to notice, that all the ethnic groups have their peak in California except for the Afro-Americans who have their highest value in New York. The agglomeration areas of the American Indians seem to be California, Oklahoma, Arizona, and New Mexiko.

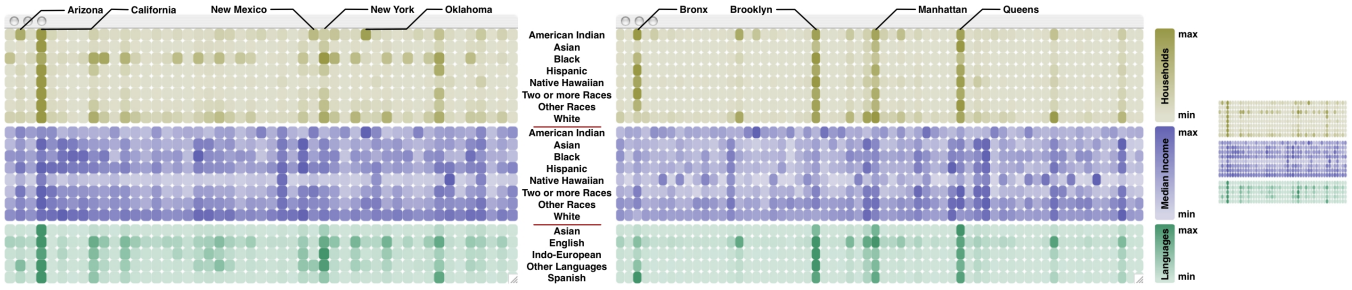
To see the proportion of each ethnic group the column-by-column normalization is the best one to use. In fig. 1c the District-of-Columbia immediately sticks out. It is the only state in which the number of Afro-Americans is higher than the number of Caucasians[2]. Regarding the median income, in every state Caucasians tend to be better situated than any other ethnic group. Furthermore,

[1]We avoided direct labelling because there are states with extensive numbers of counties (e.g., Texas). Columns in fig. 1 (left and right) are the states in alphabetical order excluding Alaska, Puerto Rico, and Hawaii but including District of Columbia and Rhode Island. Columns in fig. 1 (middle) are the counties of the state New York in alphabetical order. During interactive exploration tooltips could be used to weaken the problem of difficult orientation. To ease the interpretation of the screenshots in this paper we manually added labels for the rows and the addressed columns.

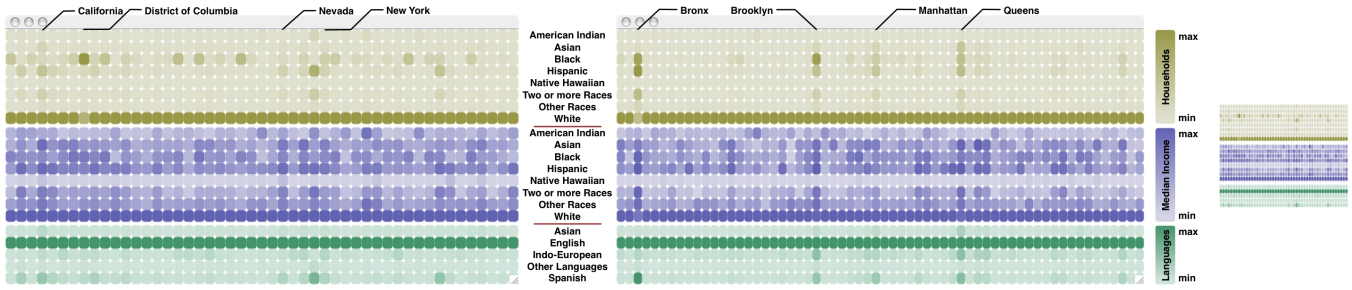
[2]Caucasian in this context is equivalent to the ethnic group "White not Hispanic or Latino" of the PUMS data set



(a) Normalization strategy based on whole matrix



(b) Row-by-row normalization strategy



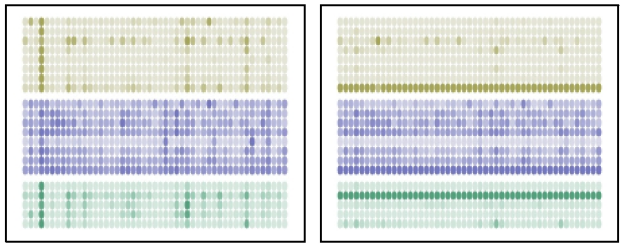
(c) Column-by-column normalization strategy

Figure 1: Visualization of the local distribution of major ethnic groups, their income and the regionally spoken languages. Geographical units are represented by columns, the data for the categories such as household, income, and language data by rows. *Left*: state level, *middle*: county level for state New York, *right*: again state level, but with an iPod-resolution of 220x176 pixel (in comparison to the other screenshots having a resolution of approx. 800x400 pixel).

a difference concerning the spread of the income between the different ethnic groups can be seen. Whereas for example in Nevada, California, or New York the differences in income are rather moderate, some other states show significant differences. Unsurprisingly, English is the main language in every state. But it is also immediately obvious that there are some states in which the ratio of Spanish speaking people is also quite high.

In fig. 1 (middle) we drilled down the state New York to see the distribution on county level. Interesting observations are for example: Fig. 1b shows that one of the main settling places for the Afro-Americans is the county Brooklyn. In fig. 1c that is based on a column-by-column normalization we can see immediately that they are also one of the major ethnic groups in this county. We can also see in fig. 1c that the Bronx is the only county in which the Caucasians are not among the major ethnic groups. Fig. 1a shows that those Caucasians that settle there also tend to have a lower income than the ones in other counties. In fig. 1c it can also be clearly seen that the proportion of Spanish speaking people in the Bronx is higher than in any other county. Furthermore, Bronx, Brooklyn, and Queens are the counties with the highest proportion of Non-English speaking people.

Finally fig. 2 shows the images of figures 1b and 1c (left) in the iPod-resolution of 220 x 176 pixel. Even with such a low resolution



(a) (b)

Figure 2: Figures 1b and 1c (left) in iPod-resolution (220 x 176 pixel) and iPod screen size (4th generation).

patterns can be perceived and the data can be explored.

References

[1] J. Chen, D. Guo, and A. M. MacEachren. Space-time-attribute analysis and visualization of u.s. company data. In *IEEE Information Visualization 2005 Proceedings Compendium*, pages 55–56, 2005.

[2] M. C. Hao, D. A. Keim, U. Dayal, and J. Schneidewind. Visual map: A visualization technique for comparison of large multi-attribute time series data. HP internal report.