Mr. Spearman or how to explore changes in trends

Last week I was with my good friend Johannes on the phone. He just became a father and called her adorable daughter "Franziska". It made me think about the meaning of the name... Franziska you could tell from the South of Germany -Bavaria- or even Austria... and a quite traditional name as well.

So names can reveal a lot about the contextual situations of a person or a time frame. Even with the proper data set you could potentially tell which periods of time follow the same naming patterns. I starting searching for baby names lists with a popularity ranking and I came to the Social Security Administration where you can pull male and female names ranked by popularity from the 1880 until last year -BTW, sometimes you cannot reach the site with a non-USA IP-Address, so you might want to have a look at proxy-like solutions like ZenMate or Hola.

Well, I managed to -obviously programmatically :)- download all the years available from the US census. I'm not providing the downloaded data, because I don't want to potentially infringe any copywriting or whatsoever law, so I just point you guys to the source. The data looked promising but I wasn't quite sure it would be possible to detect trend changes. My reasoning started like:

- Each year is a sorted list of names.
- The list from a particular year to the next or previous one is not likely to change much.
- Conversely, a "bigger-than-expected" change in the list from a year to the next one might reveal a trend interruption or a trend change

So what I need is a way of measuring how similar two ranked lists are... And good news! this metric exists and is provided by the Spearman's Correlation Coefficient for Ranks (SCCR) -BTW, the Kendall correlation coefficient does the job as well-.
For each year, I computed the SCCR value with all other years split by gender. I created a scattered plot where the size and the transparency of each point is determined by the SCCR value. Intuitively if we have a look at the diagonal, bigger opaque points together form a cluster where the trend persists and places with almost no color in this diagonal represent trend interruption.

- The average trend duration is something around 5 years for male and a bit less for female.
- The female names are more likely to change from a year to the next one, as we see less consistency over time.
- In the 19th century, people used to stick more to the baby names -see the big cluster before 1910-
- The male baby names between 1985 and 1997 didn't vary much -long lasting trend here-
- It's very unlikely, that years distant in time follow the same pattern -almost not even a single case-

Just filtering out years pairs with SCCR below 0.30 produces a picture where mostly consecutive years stay -diagonal-, while the other non-consecutive ones just disappear -with some punctual exceptions-.

Obviously, the action is in the diagonal! If we just focus on the SCCR(year i, year i+1) and map this value to 0 if it falls under a threshold or to another number otherwise, we can produce a clearer picture of the trend changes. In following picture I've played with different thresholds, which allows for understanding trend changes with different resolutions (as you can see the lower we set the threshold, the longer the trends last for).

Obviously, more sophisticated approaches can be taken, such as us the Twitter's approach for breakout detection or time series change point detection packages (good ones are CPM, BCP or ECP)... but I love simplicity and for this post the threshold mapping does a decent job.

On a separate note, the kind of data we have here allows for running Density Based Cluster Analysis - I particularly like DBSCAN or one of the improved versions of it-. I encourage the reader to get the data I'm providing below and give it a play... Please share your experience if you happen to :)

The following R code snippet computes the year-to-year SCCR for both gender values. I've exported the resulting data as csv your those who want to play with it and were too lazy to download the yearly names rank: year-to-year-sccr

```
# Year over Year Spearman generation all.years
```