Name Trend Prediction
Data Project Report for EECS 564

Yu Wang

***Abstract*—Names given by parents are more than just an identification, but also reflect the fashion of that time. As expected, new names keep springing out, some names are long lasting and some names may be more and more desolate in the future. This project will apply a Matched-Filter-based slope filter to detect the trend of the popularity. For booming period, an Epidemic Model for Name will be implemented to explain the data. After that, a model for the growth of popularity of new names will be proposed based on the result from the previous two steps.**

***Index Terms*—Names for Babies, Popularity, Epidemic Model, Prediction**

[1]

I. Introduction

Undoubtedly, everyone has a name. Parents prefer to choose names they heard about before like James and Mary, rather than random generated name like Akdbsts. In this project, we believe the spreading of a name can be described with the same model for virus, which is often called Epidemic Model[3]. Intuitively, when one couple of pioneer parents give a lovely name for their baby, their neighbors, colleagues, relatives and everyone who can have access to this piece of information will help them spread this new name. If some people give this name to their babies afterwards, a new round of spreading begins. This procedure is similar to the one that one person in the office catches cold, and several days later, many other people begins to cough. Epidemic Model is reasonable to fit this procedure. In this project, a modified model called Epidemic Model for Name(EMN) will be proposed.

However, EMN will show its power only for "spreading procedure". According to the dataset, most names are not undergoing purely spreading or fading procedure, but the combination of both or even more complex ones. So, the first step is to tag the trend with "increasing" and "decreasing", and focus on "increasing" part. The most common and naive idea to split the whole trend from 1910 to 2015 for this problem is by examining the "slope" of the curve. However, because of the existence of noise, "slope method" is not that robust. In this project, a Matched-Filter-based slope detector will be implemented. Matched-Filter[4] is the optimal solution for signal detection when the noise is AWGN(Additive White Gaussian Noise). With the help of the new detector, we can split the trend of each name more properly.

The power of Epidemic model is not limited only to fit the trend of names, but also predict the future for them. In the end, this report will also predict the future states from some names, and will propose a most likely model for the growth of a newly come up name.

A. Dataset

The dataset is given by the Social Security Administration[2]. The file can be downloaded through the link [1]. This dataset stores names' usage information for each 50 states from 1910 to 2015. To safeguard privacy, names used by less than five people in one state one certain year will not show up in the file. Each record contains state, gender, year, name and occurrence . The following record is an example of one instance in file "AK.TXT"

$$\underbrace{AK}_{State},\ \underbrace{F}_{Gender}\ ,\ \underbrace{1910}_{Year},\ \underbrace{Mary,}_{Name}\ \underbrace{14}_{\#\ of\ users}$$

The above record means that, in 1910 Alaska, 14 female babies were given name "Mary".

II. Problem Formulation & Models

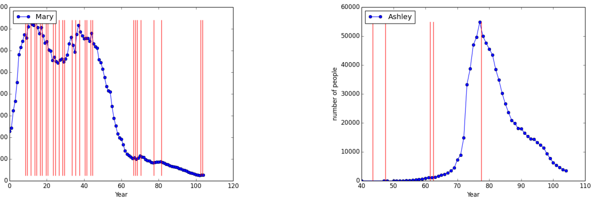
By the description of introduction in section I, three targets will be reached one by one in this project.

A. Trend Detection

The goal for Trend Detection step is to tell whether this point on the curve is in increasing or decreasing period. We should tell whether the slope for this point is greater or less than 0.

Denote the function of the curve for the name to be $y = f(t)$. For time t', we need to tell whether $k' = \frac{df(t)}{dt}|_{t=t'} \lesseqgtr 0$.

In reality, the above method is prone to noise. In other word, the slope k' may fluctuate round 0 too often.



(a) Mary

(b) Ashley

Fig. 1: Performance for Common Slope Detector

Fig.1(a) and 1(b) are two typical examples of the situation described above. Red line stands on places where the monotonicity flips. For the name "Ashley", we know that there is

[1] Code can be download through https://github.com/qizhu8/Name_Trend_Prediction

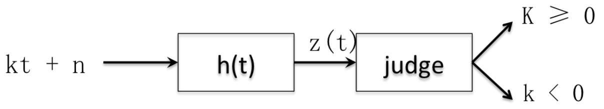


Fig. 2: Diagram for Matched-Filter-based Slope Detector

roughly one peak, and the trend is firstly going up and then going down, that simple. Because of the existence of noise, the incline part is not that smooth. As a result, many "False" detections are made. So do the curve for "Mary".

In order to make the slope detector more robust to noise, we would like to take into more points to make the decision. Then we think of the idea for Matched-Filter[4].

In this model, we assume the noise is zero mean Additive White Gaussian Noise (AWGN), with power σ^2. Also, $k \geq 0$ and $k < 0$ are equally likely. Suppose the true signal $x(t)$ is a linear function with constant slope k. Then the expression of the curve would be:

$$y(t) = \underbrace{k \cdot t}_{x(t)} + b + n \quad n \sim \mathcal{N}(0, \sigma^2), t \in \{1, 2, \cdots, m\} \tag{1}$$

Our target is to tell whether k is greater or less than 0. Then we think about the following model for detection. The hypothesis test for this problem is:

$$\begin{cases} \mathcal{H}_0 : k \geq 0 \\ \mathcal{H}_1 : k < 0 \end{cases}$$
$$p(\mathcal{H}_0) = p(\mathcal{H}_1) = 1/2$$

Denote $h(t)$ to be the patten we will use, then t_0 is the length of patten we are interested in. If $t_0 = 3$, it means we will make our decision based only on the point itself and previous, afterward one point. For $\mathcal{H}_0$, $k \geq 0$, then $z(t)$ is:

$$\begin{align} z(t|\mathcal{H}_0) &= \sum_{t=0}^{t_0-1} h(t)\big((t_0 - t)|k| + bn(t)\big) \tag{2} \\ &= |k| \sum_{t=0}^{t_0-1} h(t)(t_0 - t) + \sum_{t=0}^{t_0-1} h(t)(b + n(t)) \tag{3} \end{align}$$

For $\mathcal{H}_1$, $k < 0$, then $z(t)$ is:

$$\begin{align} z(t|\mathcal{H}_1) &= \sum_{t=0}^{t_0-1} h(t)\big(-(t_0 - t)|k| + b + n(t)\big) \tag{4} \\ &= -|k| \sum_{t=0}^{t_0-1} h(t)(t_0 - t) + \sum_{t=0}^{t_0-1} h(t)(b + n(t)) \tag{5} \end{align}$$

$z(t|\mathcal{H}_0)$ and $z(t|\mathcal{H}_1)$ are not the likelihood, but the power of output at time t_0. For the output, the one with larger Signal Noise Ratio(SNR) will win. So we get to the following expression:

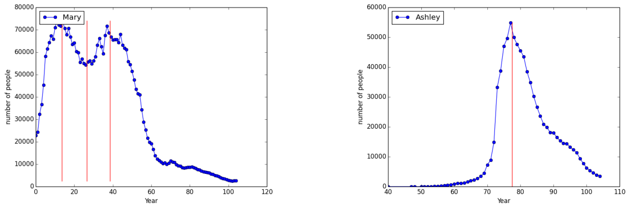
$$z(t|\mathcal{H}_0) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\lessgtr}} z(t|\mathcal{H}_1) \tag{6}$$

According to 3 and 5, we know that the only difference between them is the sign for k. We can choose $h(t)$ such that

$$\begin{align} & z(t|\mathcal{H}_0) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\lessgtr}} z(t|\mathcal{H}_1) \tag{7} \\ \Rightarrow & |k| \sum_{t=0}^{t_0-1} h(t)(t_0 - t) + \sum_{t=0}^{t_0-1} h(t)n(t) \nonumber \\ & \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\lessgtr}} -|k| \sum_{t=0}^{t_0-1} h(t)(t_0 - t) + \sum_{t=0}^{t_0-1} h(t)n(t) \tag{8} \\ \Rightarrow & \sum_{t=0}^{t_0-1} h(t)(t_0 - t) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\lessgtr}} - \sum_{t=0}^{t_0-1} h(t)(t_0 - t) \tag{9} \\ \Rightarrow & \sum_{t=0}^{t_0-1} h(t)(t_0 - t) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\lessgtr}} 0 \tag{10} \end{align}$$

From expression 10, we know that any $h(t)$ satisfies the inequality will work. So in this project, we choose $h(t) = t - t_0/2$. So, if choose $t_0 = 9$, then $h(t)$ is a sequence like $\{4.5, 3.5, 2.5, 1.5, 0.5, -0.5, -1.5, -2.5, -3.5, -4.5\}$.

So to tell whether $f(t')$ is in increasing mode, we will apply convolution to get the value after applying the patten.



(a) Mary

(b) Ashley

Fig. 3: Performance for Matched-Filter-based Slope Detector

Fig.3 shows the result for Matched-Filter-based Slope Detector for name "Mary" and "Ashley". Obviously, the performance is much better then the naive method.

B. Epidemic Model for Name

Epidemic Model for Name shares the same idea from SIR model [3] in 1994. We think the spread of a new name is similar to the epidemic of virus. After the Trend Detection II-A step, the curve of certain name from year 1910 to 2015 will be split to "incline" and "decline" periods. This step is to fit the parameter for the "incline" part's model.

For SIR model, there are "Susceptibles"(S), "Infectives"(I) and "Recovered with immunity"(R).

- $S(t)$ is used to represent the number of individuals not yet infected with the disease at time t.
- $I(t)$ denotes the number of individuals who have been infected with the disease.
- $R(t)$ is the compartment used for those individuals who have been infected and then removed from the disease.

- γ denotes the probability a susceptible get infected.
- N is the total number of population, often regarded as a fixed number.

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{11}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \tag{12}$$

$$\frac{dR}{dt} = \gamma I \tag{13}$$

We modified the model and get to Epidemic Model for Name.

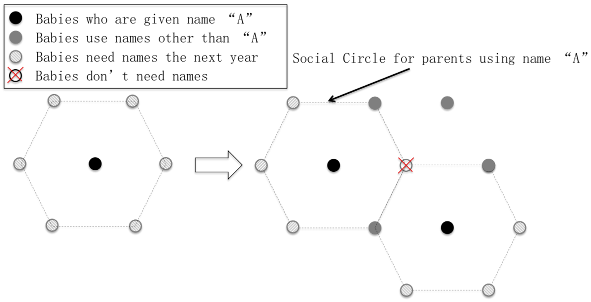

● Babies who are given name "A"
● Babies use names other than "A"
○ Babies need names the next year
☒ Babies don' t need names
Social Circle for parents using name "A"

Fig. 4: Epidemic Model for Name

When one couple of parents first come up with a very lovely name, denote as name i for their baby, they may not conceal this information to themselves, but to spread to everyone they may known. The baby who are lucky enough to get this name is like the black point in Fig. 4. When friends, relatives and colleagues of the parents know this name, they may have a certain probability p_i to give name i to their babies. Those babies who need name for the next year are marked as hatched line points. When their babies use name i, new round of spreading starts. Of course, the probability of not using name i is $1 - p_i$. In often the case, p_i is small. So many babies may have gotten name rather than i. Gray points stand for babies use other names. Needless to say, the value of p_i largely depends on how lovely name i can be. But as a result of this step, p_i for different names are not randomly scattered. Clearly, this model can only explain how one name becomes popular year after year. But this is enough for the prediction part for new names.

Now try to use mathematic language to describe this model more precisely.

At year t, the number of babies trying name i is $y_{i,t}$. For the next year, babies given the same name is $y_{i,t+1}$. If the probability of using name i for one baby is $p_{i,t}$ and the total number of newborn babies in year $t+1$ is C_{t+1}, then we have:

$$\frac{\partial y_{i,t+1}}{\partial y_i} = p_{i,t} C_{t+1} \tag{14}$$

$$\Rightarrow y_{i,t+1} = p_{i,t} C_{t+1} y_i \tag{15}$$

$$\Rightarrow y_{i,t+1} = y_1 \prod_{k=1}^{t} p_k C_{k+1} \tag{16}$$

In Eq.16, C_t can be calculated by compute the sum of all the occurrences of all names in year t, $y_{i,t}$ is the usage on name i in year t, can also be got from the dataset. The only task remains is the estimation of $p_{i,t+1}$. To simplify the model, and also based on the experimental result, we can use $\bar{p}_i = \frac{1}{t}\sum_{k=1}^{t} p_k$ to approximate $\{p_{i,k}\}_{k=1}^{t}$. So

$$p_{i,t} = \frac{y_{i,t+1}}{y_i \cdot C_{t+1}} \tag{17}$$

$$p_i = \frac{1}{t}\sum_{k=1}^{t} p_k \tag{18}$$

C. Spreading Model for New Names

Based on the assumptions and EMN model, we can build Spreading Model to predict trends for new names.

Similar to the model expressed in subsection II-B, suppose in year t, a new name comes up with population y_t, then for year $t'(t' > t)$, we will have

$$y_{t'} = y_t \cdot \prod_{k=t+1}^{t'} C_k \cdot \hat{p} \tag{19}$$

The $\hat{p}$ is the most likely p among the estimation result for the previous subsection II-B.

$$\hat{p} = \max_p p\{|\{p_i = p\}|, p_i \in P\} \Leftrightarrow \hat{p} = MAP(p_i) \tag{20}$$

Note that, P is the set for all the estimated p_i, and $|\cdot|$ is the operator to get the number of elements in a set.

III. Method

Since the size of data is too large, there are 5647426 records in all, MySQL[5] database is implemented to store the data. The processing strategy is like below:

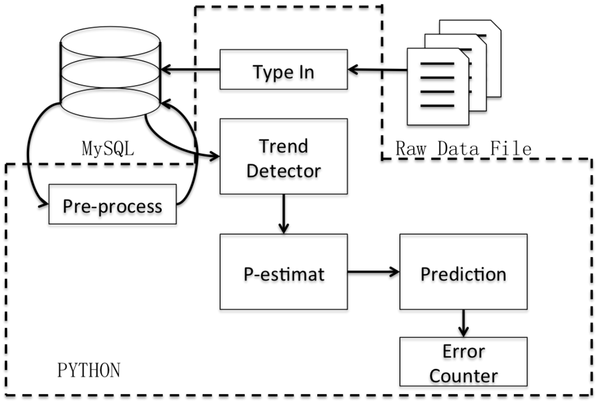
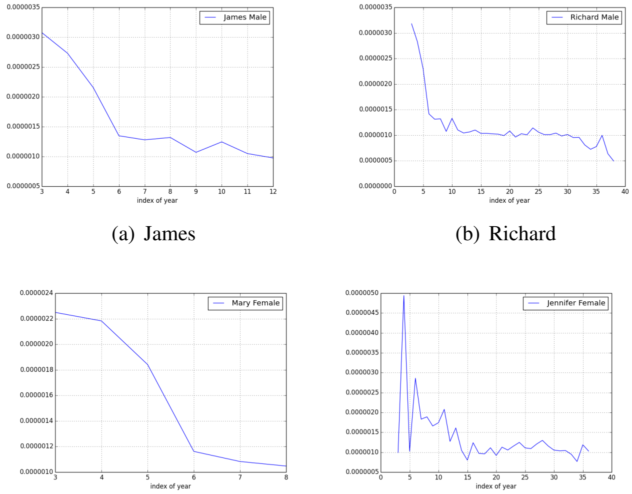


Fig. 5: Implementation Diagram

1) Type in: applying python script to type the datafile to the database;
2) Pre-process: Combine data from different states, and sort names by their occurrence;
3) Trend Detection: Apply the Matched-Filter based trend detector to get the incline part of the curve for each name;
4) P-estimation: Use the Epidemic Model for Name to estimate p for each name based on its incline period data;
5) Prediction and Counting Error: Treat the last several years (from 1 to 5) as unknown, and calculate the prediction error for the model.

IV. Results

The predicted figure for $p_{i,t}$ is like



(a) James

(b) Richard

(c) Mary

(d) Jennifer

Fig. 6: Estimated P for Different Names

The $p_{i,t}$ is not scattered randomly, but has a relatively stable value. So, using mean to represent the sequence of $\{p_{i,t}\}_{t=1}^{t'}$ is reasonable.

Based on the model produced in section II-B, the prediction error for female names is like Fig.7

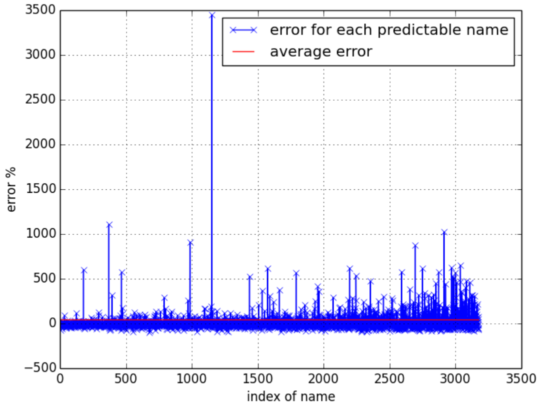


Fig. 7: Prediction Error for Female Names

The patten length is set to be 9, and only predict one year. From the figure, except we know that some names have very large errors, most predictions have error within 5%.

The estimated error versus number of year to predict is like Fig. 8.

So, the prediction for one or two years is more accurate than longer. But why the error for prediction of four years is the lowest is unknown. It may be the result of the law for the increasing trend of names.

Fig. 9 is the pdf for the estimated p. According to eq. 20, p for female names and male names are in Table. I.

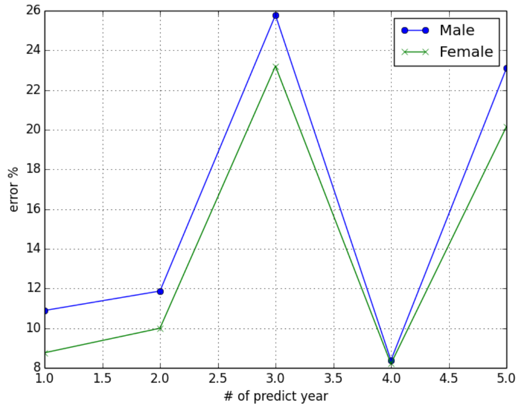


Fig. 8: Estimated Error vs Number of Year to Predict

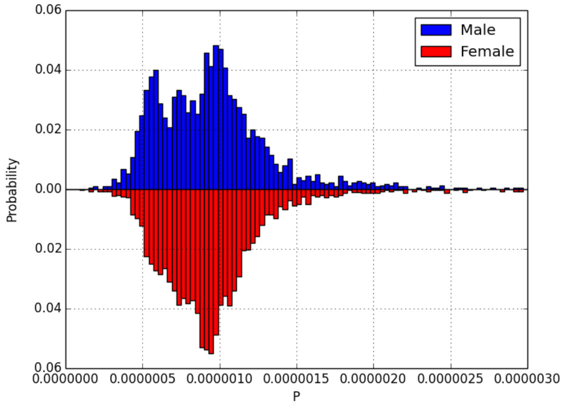


Fig. 9: pdf for Estimated P

References

[1] Social Security Administration. namesbystate.zip, 2015.

[2] Social Security Administration. Social security administration, 2016.

[3] Linda JS Allen. Some discrete-time si, sir, and sis epidemic models. *Mathematical biosciences*, 124(1):83–105, 1994.

[4] John A Gubner. *Probability and random processes for electrical and computer engineers*.

[5] AB MySQL. Mysql, 2001.

	male	female
p	9.6×10^{-7}	9.3×10^{-7}

TABLE I: P for Prediction Model of New Names