# Extracting useful information from the Google Ngram dataset: A general method to take the growth of the scientific literature into account

GTDStat

January 2016

# 1 Introduction

Recent years have seen the birth of a powerful tool for companies and scientists: the Google Ngram dataset, built from millions of digitized books. It can be, and has been, used to learn about past and present trends in the use of words over the years. This is an invaluable asset for scientists studying cultures and societies as it allows them to look for trends in the use of certain words, but also to determine when such or such concept or idea has been expressed more extensively. It is, in a way, a window on History.

The dataset is as interesting from a business perspective. There are several examples of company names whose use in the literature is strongly correlated with their stock market value. However, one will agree that the dataset's value mostly comes from its potential use in marketing. The choice of words has a major impact on the success of a marketing campaign. An analysis of the Google Ngram dataset can validate or even suggest the choice of certain words, for example choosing between synonyms like the words "smart", "clever" and "ingenious". It can also be used to predict the next buzzwords in order to improve marketing on social media. Another application is the study of the success of previous campaigns by looking at how often the company name occurs with respect to its competitors in the dataset at specific points in time.[1]

The Google Ngram dataset is a gift for scientists and companies, but it has to be used with a lot of care. False conclusions can easily be drawn from a naïve analysis of the data. It contains only a limited number of variables and that makes it difficult to use it to its full potential. The data is described in more details in the next two subsections. It is explained how the the simple nature of the dataset makes it hard to extract valuable information from it. The problem tackled in this work is described in the last subsection of the introduction. The rest of the work is divided into two parts: a presentation of the proposed method with a detailed example; and concluding remarks.

## Data

The Google Ngrams dataset is divided into different languages. In this work, the focus is on the "American English" dataset. For this corpus, there are five distinct datasets: the 1, 2, 3, 4 and 5-grams, all coming from a vast collection of digitized writings from various years.[2]

A $k$-gram is a sequence of items separated by $k-1$ spaces. Very roughly speaking, the 1-grams contained in a book are the words in it; the 2-grams are the consecutive pairs of words in it; etc., but they can sometimes be formed by sequences of items that are not actually words like the 2-gram ". The".

The datasets contain only 5 variables: *gram*; *year*; *occurrences*; *pages*; and *books*. Let $g_i$ be the $i$th gram of the list and $y_j$ the year of interest, respectively standing for the variables *gram* and *year*, then let us denote[3]

- $n_{ij}^{(o)}$ for *occurrences*: the total number of occurrences for the gram $g_i$ in the year $y_j$;
- $n_{ij}^{(b)}$ for *books*: the total number of books in which the gram $g_i$ appeared in the year $y_j$.

[1] See [2] for some examples of applications mentioned in this paragraph involving Microsoft and Goldman Sachs.
[2] For more information on the datasets, we refer the reader to the introduction in [1].
[3] The *pages* variable is omitted since it is not used in this document.

For example, in $y_j = 1945$, the gram $g_i :=$"data" has appeared in $n_{ij}^{(b)} = 2392$ books, on exactly $n_{ij}^{(p)} = 25592$ pages and a total of $n_{ij}^{(o)} = 35203$ times. As one might rightfully expect, the datasets are quite big. It does not take that many books to find an impressive number of different grams. Moreover, the *years* variable takes values from 1600 to 2008. For example, the 1-gram dataset contains $291,639,822$ rows and the 2-gram set contains almost $4 \times 10^9$ rows. Not surprisingly, the size of the datasets increases with the size of the grams considered, since that entails more possible combinations.

## Data validation

There is not much data cleaning to execute. One should be aware that some have raised doubts on the validity of the data for certain entries. These doubts come from the fact that some errors could have been introduced in the digitizing process. For example in the older books, the letter *s* is often confused with the letter *f* because of their calligraphic resemblance.[4] However, since only a limited number of grams are chosen for a given analysis, it is only necessary to investigate the behaviour of the variables for those specific entries. If the chosen grams have not occurred in a given year, then the corresponding rows are missing and rows with only 0 have to be added for these particular grams and year.

A crucial variable that is not included in the dataset is the book in which the gram has been found, i.e. a *book id* variable. Such a variable would allow much more in depth analyses. This would also provide the number of books used to construct the various datasets ($k$-grams, $k = 1, 2, 3, 4, 5$). A simple analysis of the data shows that the different datasets have not been constructed from the same collection of writings. The size of the collection used tends to grow with the size of the grams collected.

## Problem

From the structure of the dataset, it is clear that its main value is to allow the study of trends over time. However, the simple nature of the dataset adds great difficulty to its use. The *book id* variable mentioned previously could be introduced in terms of book titles or ISBN. This would allow the analyst to approximate the proportion of books coming from a particular subset (e.g. *fiction literature* or *scientific literature*.) If it was possible to partition the dataset by genre, one would realize that the relative size of each subset varies over time. This poses an important problem when one wants to study the trends over time: the results of a straightforward analysis are strongly dependent on the variation by years of the proportion of books coming from a given type of literature (see [3], for an explanation of the Simpson's paradox). Moreover, it is not even clear whether one should take for granted that these proportions reflect reality.

In summary, to perform a meaningful trend analysis, one would need not only to properly identify the types of writings he is interested in, but also to isolate the evolution of the use of the gram of interest within these subsets. The originality of this work lies in the clever approximations and estimations of the missing book type or id needed to do this using proxies and statistical models. The proposed method can be viewed as a tool that provides a basic defense against some misleading trends that can be found in the data.

[4] An appropriate Google search provides many examples of blog entries on the matter.

# 2 Analysis

## Data transformation

A crucial choice to make is which variable to analyse. Various reasons can lead an analyst to use the brute number of occurrences of a given word over the years, but one might instead want to analyse the use of a particular gram by the number of books in which it appeared. The focus is put on an analysis with the *books* variable in this work. The first step is to transform the brute data into proportions. This is where the numbers of digitized books (for each year) are needed. Although not provided, these values can be approximated. For the 1-gram database, the proxies used are the maximum of the *books* variable for each year. For the 2-gram database, the gram "of the" was chosen after some graphical experiments, since the computations for the maximum (over the *books* variable) for each year are extremely lengthy.[5] For the studied gram $g_{i_0}$, let

$$p_j := \frac{n_{i_0 j}^{(b)}}{\max_i n_{ij}^{(b)}} \approx \frac{n_{i_0 j}^{(b)}}{\#\text{ books for the year } y_j} \tag{1}$$

be the year-adjusted value of the variable $n_{i_0 j}^{(b)}$.[6] For example, if we look at the gram "competition", we get that it has appeared in 12 books in 1967 and the maximum for the variable books in 1967 in the 1-gram dataset is 100. Its normalized *books* variable is then 12/100.

To simultaneously model the changes in proportion of the selected gram $g_{i_0}$ with respect to two complementary subsets of the digitized books, noted $A$ and $A^c$, an ingenious way to proceed is to choose a gram that is known to be contained in almost all books of $A$ and almost none of $A^c$, or vice versa. Let this gram be denoted $g_{i_A}$. Again, let

$$q_j := \frac{n_{i_A j}^{(b)}}{\max_i n_{ij}^{(b)}} \approx \frac{n_{i_A j}^{(b)}}{\#\text{ books in } y_j} \tag{2}$$

be the normalized version of books for $g_{i_A}$.[7] Finally, for the sake of clarity, let $t_j := y_j - y_0$ for $i = 1, \ldots, T$, where $[y_0, y_T]$ is the interval of years for which the use of $g_{i_0}$ is studied.

It is important to stress that it is not the Google Ngram dataset that is partitioned. The subsets are only theoretical. However, it is still possible to approximate the evolution of the size of these subsets over time.

## Statistical model

The proposed model is a linear regression but with autocorrelated errors. It is easier to interpret under this form:

$$p_j \;=\; \beta_{11}\, q_j + \beta_{12}\, (1 - q_j) + \beta_{21}\, q_j\, t_j + \beta_{22}\, (1 - q_j)\, t_j + \epsilon_j \quad j = 0, \ldots, T,$$

[5] Computations have been terminated shortly after 30 minutes. It might well be possible to obtain the desired values in a reasonable time.

[6] This is the version for the grams of size one, as explained in the paragraph.

[7] Note that the normalization constants are not the same for two grams of different sizes. An additional index should be used to indicate from which dataset the values come from, but it is omitted for the sake of clarity. Refer to Appendix 4.2 for an alternative normalization when the *occurrences* variable is used.

where $\beta_{1\ell}$ represents the proportion of books in subset $\ell$ that contain $g_{i_0}$ at time $y_0$ while $\beta_{2\ell}$ represents the linear change over time of books in the same subset $\ell$. Rearranging terms, we get

$$p_j \;=\; \beta_0 + \beta_1\, q_j + \beta_2\, t_j + \beta_3\, t_j\, q_j + \epsilon_j, \quad j = 0, \ldots, T. \tag{3}$$

$\epsilon_j$ is a standard AR($p$) process, where the order $p \in \mathbb{N}$ is selected using the classical BIC criterion.

**Interpretation of the parameters**

- $\beta_0 = \beta_{12}$: the proportion of books containing $g_{i_0}$ outside the subset of the literature $A$ at year $y_0$;
- $\beta_1 + \beta_0 = \beta_{11}$: the proportion of books containing $g_{i_0}$ within the subset of the literature $A$ at year $y_0$;
- $\beta_2 = \beta_{22}$: the change in the proportion of books containing $g_{i_0}$ outside the subset of the literature $A$ from one year to the next;
- $\beta_3 + \beta_2 = \beta_{21}$: the change in the proportion of books containing $g_{i_0}$ within the subset of the literature $A$ from one year to the next.

**Extension to $m$ subsets**

The model can be extended to a partition of $m$ subsets of the literature, with

$$p_j \;=\; \beta_0 + \sum_{\ell=1}^{m-1} \beta_1^{(\ell)}\, q_{\ell j} + \beta_2\, t_j + \sum_{\ell=1}^{m-1} \beta_3^{(\ell)}\, t_j\, q_{\ell j} + \epsilon_j,$$

where $q_\ell$ is the (approximated) proportion of books coming from the $\ell$th subset. One could, for example, define the subsets *fiction*, *science* and *rest*, where the last subset would take care of the books not included in the first two. Model (4) is the particular case with $m = 2$ subsets.

## Detailed example

As a team of PhD students, we are particularly interested in the scientific literature. The subsets $A$ and $A^c$ introduced for this example are therefore *scientific* versus *non-scientific*. The gram studied is $g_{i_0}$ ="Figure", not to be confused with "figure". Figure 4.1 shows the values of *occurrences* and *books* for the two grams by years. A naïve analysis of the trends of these words is somewhat misleading. Across the 20th century, the popularity of the word "Figure" seems to increase quite sharply.

Except when at the beginning of a sentence, the word "Figure" is almost only used in scientific writings, when one wants to refer to a certain image (e.g. "Figure 5 shows..."). This implies that a change in the proportion of scientific literature over the years could obviously affect the conclusions of a straightforward analysis of the occurences of "Figure".

The proportion of books used to construct the dataset coming from the scientific literature tends to increase quickly with the years, starting from around 1950 (see [1]). The task is to jointly

analyse the evolution of the use of "Figure" within the two subsets of the literature *scientific* ($A$) and *non-scientific* ($A^c$) for the years $y_0 = 1900$ to $y_{100} = 2000$, inclusively.

To approximate the proportion of books coming from the *scientific* subset, the proxy used is the 2-gram $g_{i_A}$ :="et al". It is known to be contained in almost all scientific writings (to refer to a certain work as in "Duchesne et al, 2015") and contained in almost no non-scientific writing.

The autoregressive regression model of eq. (4) was fitted using the PROC AUTOREG procedure of SAS (maximum likelihood estimation, for more details on the procedure see Appendix 4.3). Table 1 presents the results obtained for the three popular grams inside the scientific literature: "Figure", "figure" and "model". The proportion of books containing "Figure" (resp. "figure")

Table 1: Fit of the regression model (4) with scientific words. The standard errors of the estimated coefficients are presented between parentheses. The errors $\epsilon_j$ are supposed AR($p$).

| word | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $p$ |
|---|---|---|---|---|---|
| "Figure" | 0.0749 | 0.6009 | 0.0018 | -0.0034 | 3 |
| | (0.0128) | (0.1660) | (0.0004) | (0.0015) | |
| "figure" | 0.7639 | -0.0970 | 0.0013 | -0.0002 | 1 |
| | (0.0077) | (0.1067) | (0.0003) | (0.0009) | |
| "model" | 0.4973 | -0.0180 | 0.0013 | 0.0041 | 1 |
| | (0.0246) | (0.2096) | (0.0008) | (0.0023) | |

outside the scientific literature at year 1900 ($\hat{\beta}_0$) is 7.49% (resp. 76.39%). Still at year 1900, within the scientific literature the proportion of books containing "Figure" (resp. "figure") ($\hat{\beta}_1 + \hat{\beta}_0$) is estimated as 67.58% (resp. 66.69%). As expected, this suggests that the estimated proportion of "Figure" is higher within the scientific literature. This effect is also seen with "figure", but to a lesser extent. Both linear trends for "Figure" and "figure" each are small, but not negligible. The conclusion is that the word "Figure" becomes more popular with the years, but this increase is not significantly different from the increase in the proportion of the scientific literature.

According to the Table 1, the proportion of books containing "model" outside the scientific literature at year 1900 is 49.73%, whereas the proportion of books containing "model" within the scientific literature at year 1900 is 47.93%. Linear trends for "model" are small and negligible. It is interesting to remark that the method has captured the fact that the words "model" and "figure" are also very common outside the scientific literature.

**Generalization**

What needs to be stressed here is that a statistical analysis without the subset differentiation would lead to a somewhat misleading result. It is not really that the word "Figure" has become way more popular in the second half of the century, but rather that the scientific literature has taken over. The model and the use of the proxy "et al" allow the analyst to decompose the observed use of the word "Figure" into three components: its use within the *scientific* subset; its use within the *non-scientific* subset; and the change in proportion of the two subsets. The proposed decomposition is shown in Figure 4.1. It is then obvious that what is observed is mainly explained by the third component and that no significant gain in popularity is observed. Now, the analysis has been done for the gram "Figure", but it can be done with much more interesting words, depending on the application one has in mind. In that case, it might be appropriate to define the subsets differently.

# 3 Final Comments

## Suggestions for future studies

### Partitioning

So far, the decision process for the partitioning (e.g. *scientific* vs *non-scientific*) is in no way data-driven. An improvement would be to find a way to identify these subsets automatically from the data rather than defining them *a priori*. This would entail a complex correlation analysis of the grams. For example, if almost all common words in the medical publications suddenly occur more often at a certain point in time, one might suspect that this is due to a sharp increase in proportion of the medical publications used to construct the dataset for this period. The challenge is to differentiate groups of grams whose use simply increases simultaneously from groups of grams whose increase in occurrences is due to the growth of a particular subset of the literature.

### Choice of the model

The model chosen for this work can only manage linear trends in time. The goal was to show how much accounting for the heterogeneity (in genre) of the books used to construct the dataset can lead to an improved analysis, it was not to propose a sophisticated model. There are plenty of gram behaviours that cannot be modelled adequately within this framework. One could focus on constructing a more flexible model that could take care of sudden peaks in popularity of certain words. This suggests a potential transformation of the data that could be used to model such a phenomenon. An example of such a model is presented in Appendix 4.4, where an example introduces a model that can manage piecewise linear trends. The word of interest is "vampire".

## Conclusion

In this work, it is shown that social scientists and business/marketing analysts can extract useful information from the Ngram datasets with careful analysis of the use of a given word over time. To do so, one has to clearly define the type of books he is interested in and question himself on the possible misrepresentation of the genres within the datasets. In particular, it is clear that the datasets are dominated by the scientific literature in the second half of the 20th century (especially medical publications, see [1]). The problem is that the needed information to account for this bias is not directly available.

This paper provides a careful approach in which the seemingly unavailable information, i.e. the proportion of books used to construct the datasets that come from a particular genre of writings, is approximated. A linear regression was used to demonstrate the impact of the growth in proportion of the scientific literature with the years and how it can lead to questionable conclusions. The method allows the decomposition of the use of a given word over time in multiple simultaneous trends. Its main drawback is that it might be difficult to find appropriate grams to properly define subsets. Grams like "et al" that are contained almost only in a specific subset of the literature, here the scientific literature, seem quite rare. Nevertheless, it is obvious from this work that although the observed trends in the Google Ngrams dataset have to be analysed very carefully, an ingenious analysis of the data it provides can lead to meaningful and useful discoveries.

# References


[1] Eitan Adam Pechenick, Christopher M. Danforth and Peter Sheridan Dodds, *Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution*, Plos One, 10 (**10**), 2015.

[2] Gauthier Dupont, *Language usage revealing business trends*, `dupontconsulting.wordpress.com/2012/04/02/language-usage/`, 2012.

[3] Steven A Julious and Mark A Mullee, *Confounding and Simpson's paradox*, BMJ, 309 (**6967**), 1994.


# 4 Appendices

## 4.1 Visualization

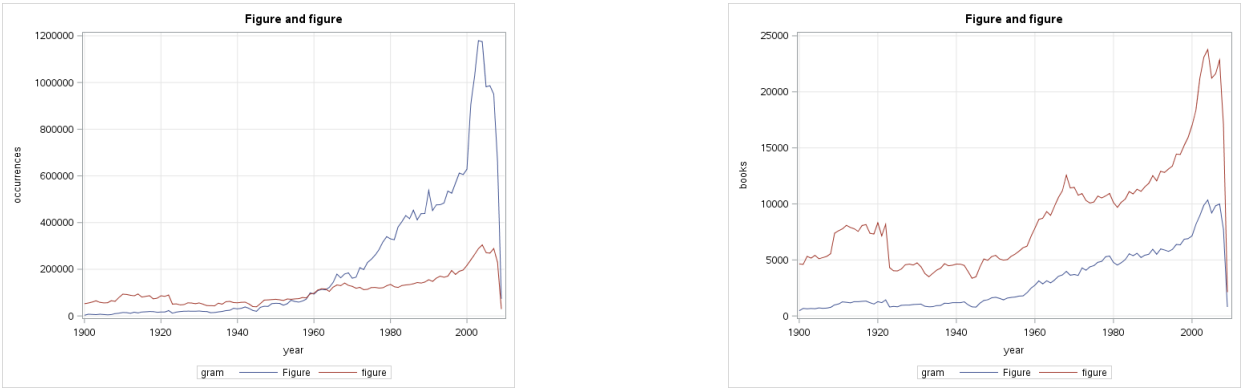


Figure 1: Variables occurrences and books for the grams "Figure" and "figure" form 1900.

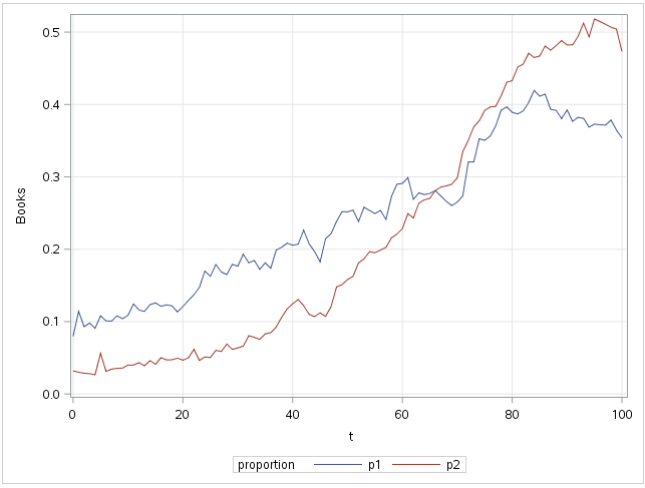


Figure 2: Proportion of the grams "Figure" and "et al" from year 1900 to year 2000.

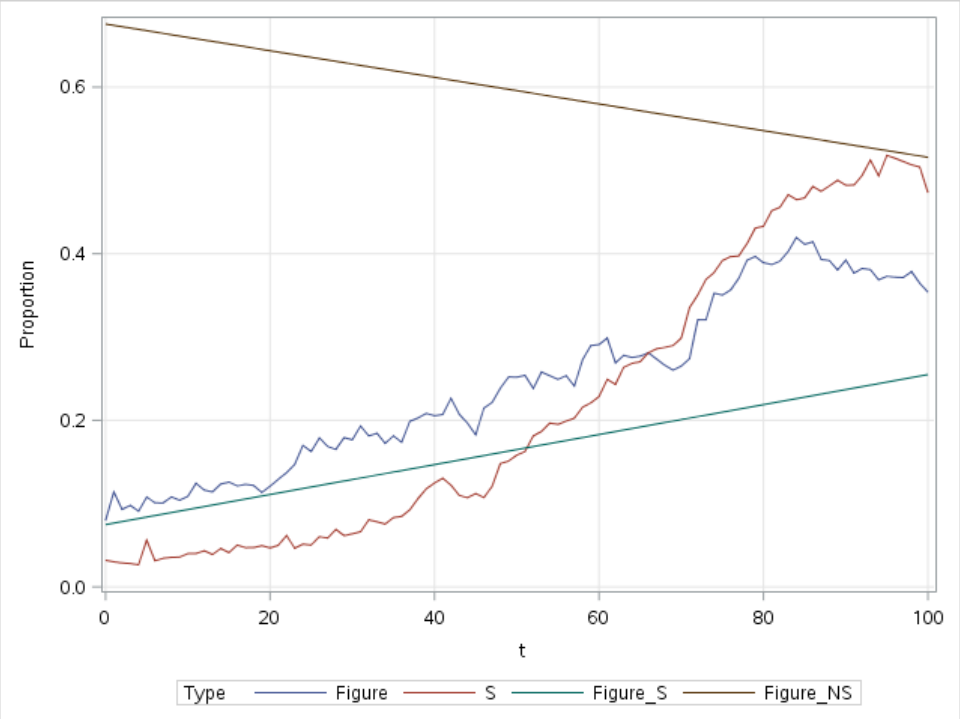


Figure 3: Figure is the proportion of books containing the gram "Figure" over time (obtained from the data). S is the proportion, by years, of scientific books used to construct the dataset (approximated with the data). Figure_S and Figure_NS are the proportions of books containing the gram "Figure" within the *scientific* and *non-scientific* subsets respectively. $t = 0$ corresponds to the year 1900. Figure is (approximately) a linear combinaison of Figure_S and Figure_NS determined by the value of S. For example, if S$\approx 1$ all along, Figure would closely follow Figure_S and similarly, if S$\approx 0$ all along, Figure would closely follow Figure_NS.

## 4.2 Normalization when the *occurrences* variable is chosen

When the *occurrences* variables is chosen instead of the books variable, the normalization should be

$$p_j := \frac{n_j^{(o)}}{\sum_i n_{ij}^{(o)}}.$$

Here, no approximation is necessary since the appropriate information is available, that is to say the total number of $k$-gram occurrences (for the appropriate $k$).

The explanatory variables should still be transformations of the $n_{ij}^{(b)}$, since they represent proportion of writings coming from a given subset. The interpretation of the coefficient becomes a bit more tricky than the interpretation using the *books* variable (see the interpretation of the coefficients of the model (4)). For example, one of the coefficients can be interpreted in the following way: the proportion of occurrences of the studied gram in the overall dataset coming from a proportion unit (in books) of a given subsets. This corresponds to the multiplication between the proportion of total occurrences coming from the subset and the proportion of occurrences of the gram of interest within that subset. Unfortunetely, these two values are not tractable and have to be approximated.

## 4.3 Likelihood of model (4)

Model (4) can be written

$$p_j \;=\; \beta_0 + \beta_1\, q_j + \beta_2\, t_j + \beta_3\, t_j\, q_j + \epsilon_j, \quad j = 0, \ldots, T,$$

where $\epsilon_j$ is a standard AR($p$) process. Again for $j = 0, \ldots, T$, let $x_j := (1, q_j, t_j, t_j q_j)$. For the particular case $p = 1$, the likelihood function of $\beta := (\beta_0, \beta_1, \beta_2, \beta_3)^\top$, $\sigma^2$ and $\phi$ is given by the product of the one-step ahead predictive densities

$$L(\beta, \sigma^2, \phi) = \frac{(1-\phi^2)^{1/2}}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=0}^{T}\left(p_j^\star - x_j^\star\beta\right)^2\right\}, \tag{4}$$

where $p_j^\star$ and $x_j^\star$ are transformed variables such that

$$\begin{aligned} p_0^\star &= (1-\phi^2)^{1/2} p_0; \\ x_0^\star &= (1-\phi^2)^{1/2} x_0; \\ p_j^\star &= p_j - \phi p_{j-1}, \quad j = 1, \ldots, T; \\ x_j^\star &= x_j - \phi x_{j-1}, \quad j = 1, \ldots, T \end{aligned}$$

and $\phi$ is the autocorrelation coefficient, that is

$$\epsilon_j = \phi\epsilon_{j-1} + u_t, \;\; u_t \sim N(0, \sigma^2), \quad j = 1, \ldots, T.$$

## 4.4 Piecewise linear slope

Piecewise linear trends over the years can be taken into account with the following model

$$p_j \;=\; \beta_0 + \beta_1\, q_j + \sum_{l=1}^{m-1} \mathbf{1}_{[t_{k_l}, t_{k_{l+1}}[}(t_j)\left(\beta_2^{(l)}\, t_j + \beta_3^{(l)}\, t_j\, q_j\right) + \epsilon_j, \tag{5}$$

where $j = 0, \ldots, 100$ and $\{t_{k_1}, \ldots, t_{k_m}\} \subset \{0, \ldots, 100\}$ defines the time intervals.

### "vampire" example

The next example is the word "vampire", which we expect to be almost non-existent in the scientific literature.

Figure 4.4 shows only one change of trend (around 1978) in the proportion ($p_j$) of the word across the 20th century, but a more accurate analysis shows a first break within the non-scientific subset at 1940. This break is masked by the simultaneous growth of the scientific literature within the datasets. This is confirmed by the 3-slopes model fitted.

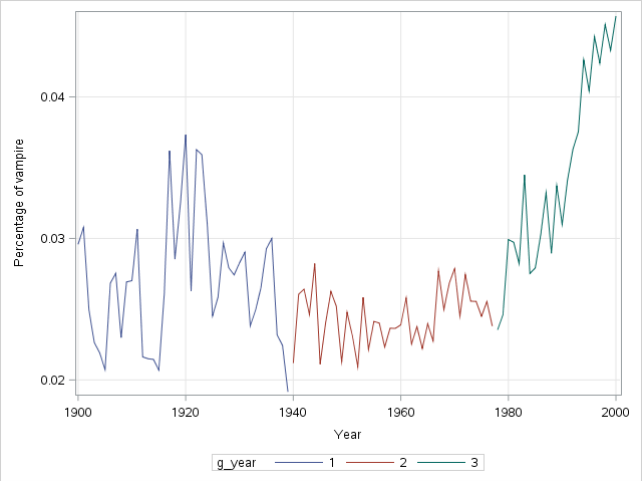


Figure 4: Pourcentage of the word "vampire" ($p_j$) decomposed as $[1900, 1940[$, $[1940, 1978[$ and $[1978, 2000[$.

Table 2: Fit of the piecewise regression model 5 with the word "vampire". The errors are considered AR(1). The standard errors of the estimated coefficents are presented between parenthesis.

| year interval | intercept | $q_j$ |
|---|---|---|
| $[1900, 2000]$ | 0.0291<br>(0.0019) | -0.1483<br>(0.0001) |

| year interval | $t_j$ | $t_j q_j$ |
|---|---|---|
| $[1900, 1940[$ | 0.0004<br>(0.0001) | -0.0016<br>(0.0016) |
| $[1940, 1978[$ | 0.00001<br>($5.8e^{-5}$) | 0.0015<br>(0.0003) |
| $[1978, 2000]$ | $-2.5e^{-5}$<br>(0.0002) | 0.0018<br>(0.0003) |

## 4.5 Main SAS functions used

### Construction of the datasets for the example of vampire

```
proc sql noprint;
        create table WORK.vampire as select * from ENGNGRAM.ENGLISH_ALL_1GRAMS
                where((gram EQ "vampire" AND year GE 1900 AND year LE 2000) );
quit;



proc sql noprint;
        create table WORK.etal as select * from ENGNGRAM.ENGLISH_ALL_2GRAMS
                where((gram EQ "et al" AND year GE 1900 AND year LE 2000) );
quit;


ods noproctitle;
```

```
ods graphics / imagemap=on;

proc means data=ENGNGRAM.ENGLISH_ALL_1GRAMS(where=(year GE 1900 AND year LE 2000))
chartype max vardef=df;
        var books occurrences;
        class year;
        output out=WORK.Max1gram max= / autoname;
run;


data work.Max;
set work.Max1gram;
if _N_=1 then delete;
run;

data work.Max;
set work.Max;
gram="MAX";
run;

proc datasets;
MODIFY Max;
RENAME books_Max=books occurrences_Max=occurrences;
run;
data work.Max;
set work.Max;
drop _TYPE_ _FREQ_;
run;

proc sql noprint;
        create table WORK.The as select * from ENGNGRAM.ENGLISH_ALL_2GRAMS
                where((gram EQ "of the" AND year GE 1900 AND year LE 2000) );
quit;

data WORK.All2;
set work.etal work.the work.vampire work.max;
run;


proc sort data=WORK.ALL2 out=WORK.SORTTempTableSorted;
        by year gram;
run;
proc transpose data=WORK.SORTTempTableSorted prefix=Books_
                out=WORK.Split3(drop=_Name_);
        var books;
```

```
        id gram;
        by year;
run;
proc delete data=WORK.SORTTempTableSorted;
run;


data work.split3;
set work.split3;
Books_Reste='Books_of the'n- 'Books_et al'n;
drop _LABEL_;
run;



proc sort data=WORK.ALL2 out=WORK.SORTTempTableSorted;
        by year gram;
run;
proc transpose data=WORK.SORTTempTableSorted prefix=occurrences_
                out=WORK.Split4(drop=_Name_);
        var occurrences;
        id gram;
        by year;
run;
proc delete data=WORK.SORTTempTableSorted;
run;
data work.split4;
set work.split4;
occurrences_Reste='occurrences_of the'n- 'occurrences_et al'n;
drop _LABEL_;
run;


data datareg3;
merge split3 split4;
by year;
run;
```

**Construction of the main variables**

```
data datareg3_books;
set datareg3;
by year;
Per_vampire = Books_vampire / Books_MAX;
Per_etal = 'Books_et al'n / 'Books_of the'n;
```

```
Per_Reste = 'Books_Reste'n / 'Books_of the'n;
DROP Books_vampire Books_MAX 'Books_et al'n 'Books_of the'n 'Books_Reste'n
occurrences_vampire occurrences_MAX 'occurrences_et al'n 'occurrences_of the'n 'occurrences_Reste'n;
run;


ods graphics / reset imagemap;

proc sgplot data=WORK.DATAREG3_BOOKS;
        ;
        series x=year y=Per_vampire / transparency=0.0 name='Series';
        xaxis grid label="Year";
        yaxis grid label="Percentage of vampire";
run;
ods graphics / reset;

data WORK.datareg3_books_tr_yr;
        set WORK.DATAREG3_BOOKS;
        tr_year=year-1900;
        if year<1940 then
        g_year=1;
        if 1940<=year<1978 then
        g_year=2;
        if year>=1978 then
        g_year=3;
run;
```

**Regression model**

```
/* Modele classique */
proc autoreg data=datareg3_books_tr_yr;
        CLASS g_year;
    model Per_vampire = tr_year*g_year / nlag=5 BACKSTEP method=ml;
run;

proc autoreg data=datareg3_books_tr_yr;
        CLASS g_year;
    model Per_vampire = Per_etal tr_year*g_year Per_etal*tr_year*g_year / nlag=5
     BACKSTEP method=ml;
run;
```