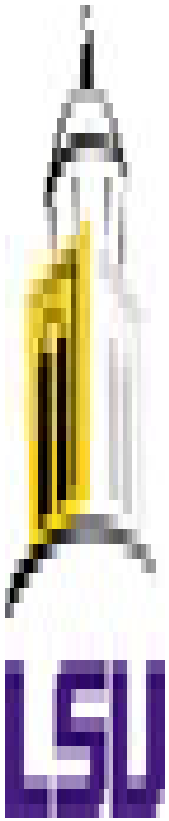
Forecasting Usability of the English Transitional Words Database



Carlos García

Seminar in the Department of Experimental Statistics

Louisiana State University A&M

February 17, 2010

Outline

- Transitional words
- English Transitional Words Database (ETWD)
  - Usage of ETWD
  - Importance of transitional words
- Forecasting problem: Page views & Visits
  - Descriptive statistics
  - Trend & Unit roots
  - ARMA model
  - Forecasting
- Conclusions and Recommendations

Transitional Words

- They can be either verbs, conjunctions, adverbs or prepositional phrases that intend to have an effect in the direction of the rhetorical objective
  - they provide a logical flow to the composition.
- They establish different relationships (links) among ideas/objects and can be placed in many places in the text :
  - between paragraphs, between sentences, within and between the parts of a sentence.
  - providing logical flow, coherence and unity to the text.
- Proficiency in reading and writing depends on how well *transitional words* are interpreted
  - Such interpretation determines message conveyance.

English Transitional Words Database ETWD

- Creation with the objective of organizing *transitional words* based on usage patterns rather than purely grammatical rules.
  - functionality that takes in the process of constructing arguments (direction).
- English skills (Listening, Reading, Writing & Speaking):
  - ESL students, universities, school districts, normal users, technical writers, etc.

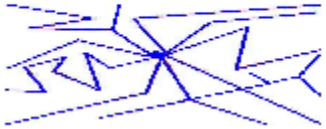
English Transitional Words Database ETWD - Windows Internet Explorer

File Edit View Favorites Tools Help

http://www.carlosignacio.com/twd/view_cat.php

English Transitional Words Database ETWD

English Transitional Words Database 2.0

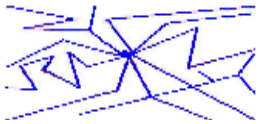


Categories

Categories available: 42 Add category

Categories	Description	Princeton	Merrian Webster
Addition	-Add statements or continue sentences that were expressed in the preceding sentence.	Princeton	Merrian Webster
Alternative	-Ideas are presented as alternatives or substitutes.	Princeton	Merrian Webster
Apposition	-Relate statements with a relationship such that one statement serves to define or modify the other.	Princeton	Merrian Webster
Cause	-Express the causes and reasons of facts.	Princeton	Merrian Webster
Chain	-Presents a series of related statements that depend on each other.	Princeton	Merrian Webster
Clarification	-Helps to restate arguments with the purpose of obtaining a better understanding.	Princeton	Merrian Webster
Combinations	-Common combinations of transitional words.	Princeton	Merrian Webster
Common Adverbs	-List of common adverbs.	Princeton	Merrian Webster
Comparison	-Ideas are shown for comparison in order to express either differences or similarities.	Princeton	Merrian Webster
Concession	-Express statements that yield or concede control of another.	Princeton	Merrian Webster
Conclusion	-Aid to make conclusions from a previous proposition or set of propositions.	Princeton	Merrian Webster

English Transitional Words Database 2.0

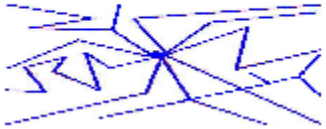


›Category: Addition

/ords available: 44 Add word

Examples	Transitional Words	Description	Traducido por Yahoo!	Princeton	Yahoo!	Dictionary	Thesaurus	Meriam-Webster	DFD
Examples	a further	-	Español	Princeton	Yahoo!	Dictionary	Thesaurus	Meriam-Webster	DFD
Examples	additionally	-	Español	Princeton	Yahoo!	Dictionary	Thesaurus	Meriam-Webster	DFD
Examples	again	-	Español	Princeton	Yahoo!	Dictionary	Thesaurus	Meriam-Webster	DFD
Examples	along with	-in conjunction with, in addition to, additionally	Español	Princeton	Yahoo!	Dictionary	Thesaurus	Meriam-Webster	DFD
Examples	also	-	Español	Princeton	Yahoo!	Dictionary	Thesaurus	Meriam-Webster	DFD
Examples	and	-	Español	Princeton	Yahoo!	Dictionary	Thesaurus	Meriam-Webster	DFD
Examples	and then	-	Español	Princeton	Yahoo!	Dictionary	Thesaurus	Meriam-Webster	DFD
Examples	another	-	Español	Princeton	Yahoo!	Dictionary	Thesaurus	Meriam-Webster	DFD
Examples	as well	-	Español	Princeton	Yahoo!	Dictionary	Thesaurus	Meriam-Webster	DFD

English Transitional Words Database 2.0



Category: Addition

Transitional Word: a further

Add Example

Examples

The act where Maria starts is a further proof that her career in TV is going to be bright.
They have sent me a further explanation of the difference between chaos and conscience.
The affirmation earlier this noon of this constitutional freedom to speak our minds is a further example of freedom and greatness of human kind.

Add Example

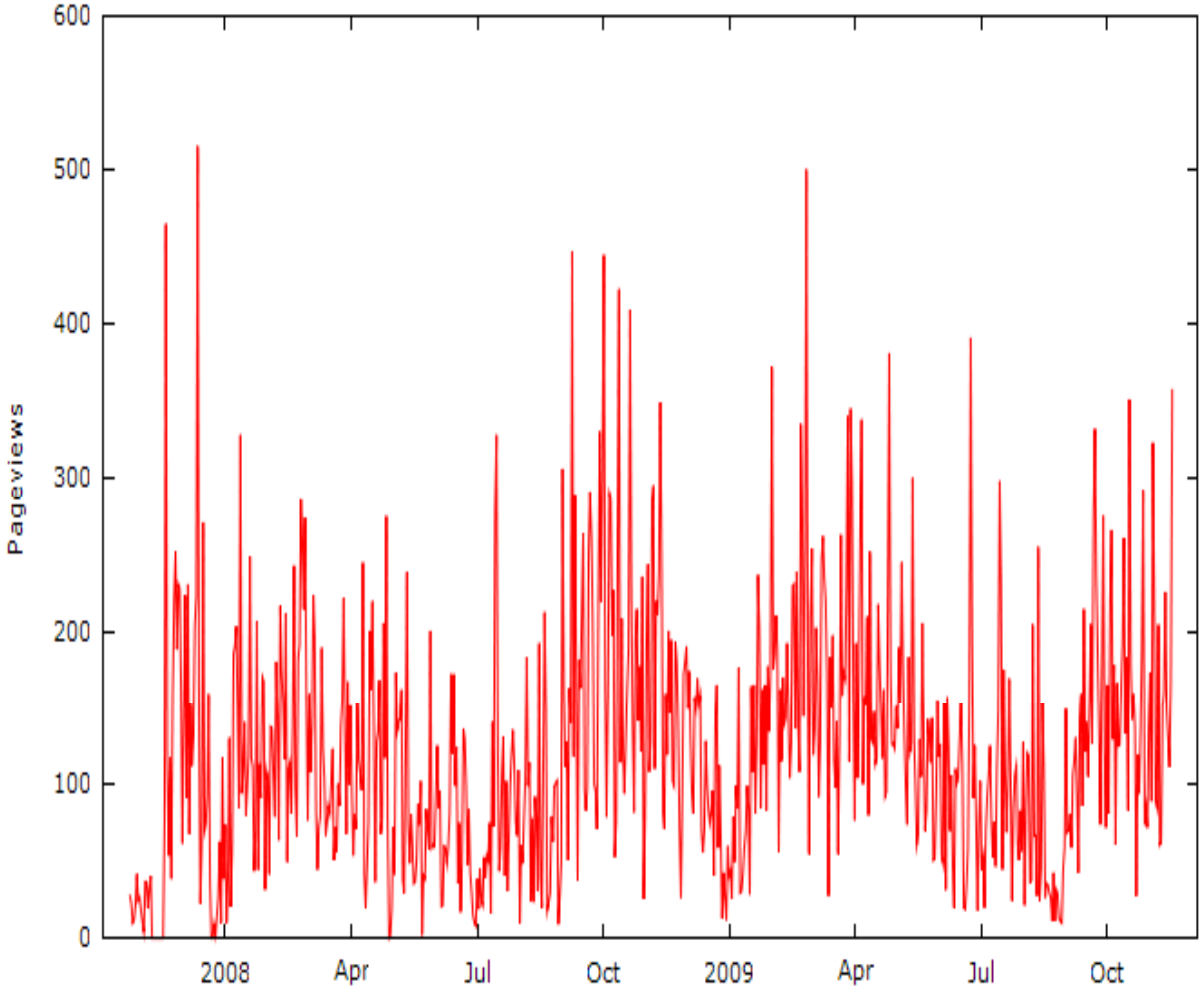
If examples are not displayed, please feel free to add!

Categories on the ETWD

- As of 03/01/10
  - Categories available: 42
  - Words available: 1261
- More common categories for relationships of:
  - Time: 103
  - Spatial: 87
  - Addition: 44
  - Causality: 44

Daily Page Views

10/24/07 - 11/18/09

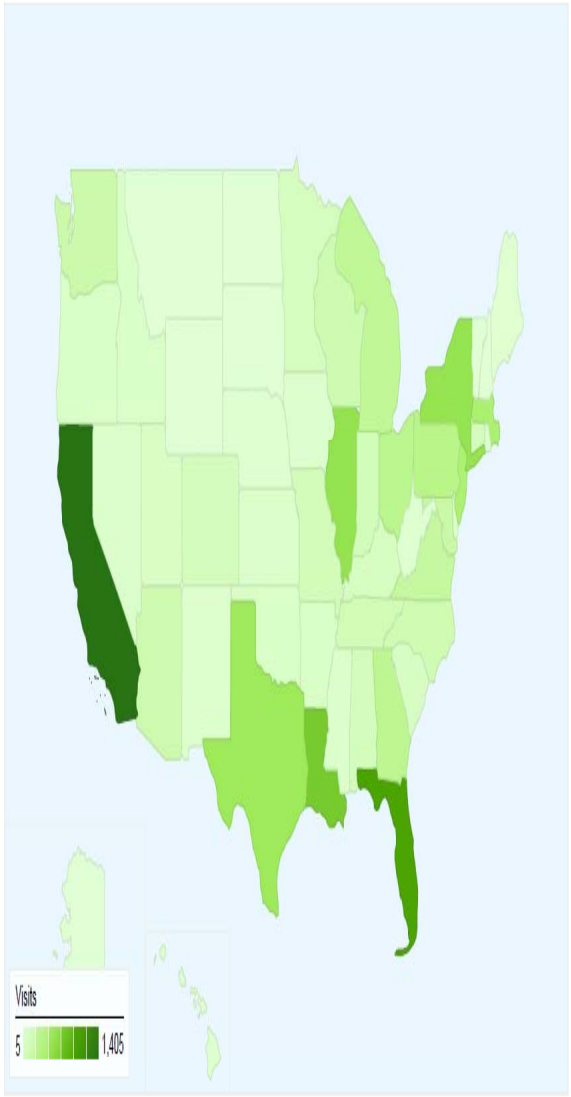


Statistic	Value
Mean	122.94
Median	111.00
Minimum	0.00000
Maximum	515.00
Standard deviation	82.831
C.V.	0.67378
Skewness	1.1384
Ex. kurtosis	1.9859

Usage > around the World

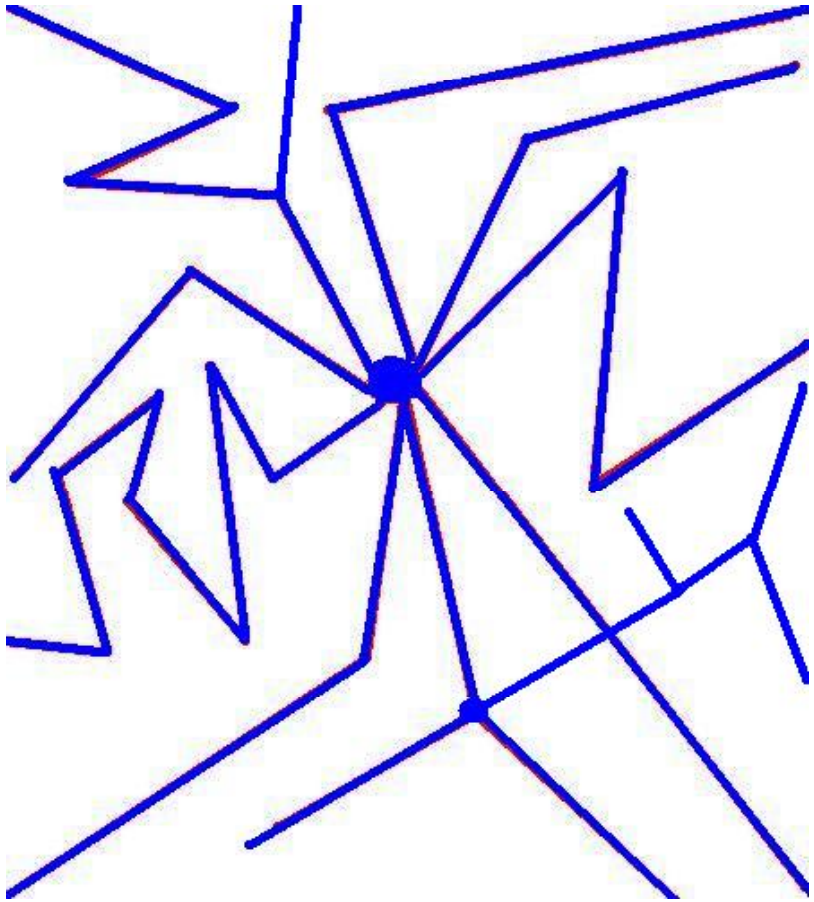
Rank	Country	Pages/visit	Time on Site	New Visitors
1	United States	6.2	0:04:47	0.7184
2	Canada	6.1	0:05:06	0.6216
3	Philippines	7.37	0:05:05	0.7491
4	South Korea	8.8	0:06:31	0.7569
5	Germany	9.16	0:06:29	0.5025
6	United Kingdom	7.39	0:04:27	0.7619
7	Australia	7.57	0:05:31	0.6236
8	China	7.46	0:06:00	0.7143
9	Taiwan	7.28	0:06:05	0.5669
10	Singapore	7.62	0:08:34	0.6098
11	Malaysia	5.77	0:04:57	0.6881
12	India	12.41	0:08:32	0.717
13	France	8.74	0:06:50	0.4021
14	Finland	7.08	0:08:05	0.253
15	Russia	5.54	0:03:20	0.3415

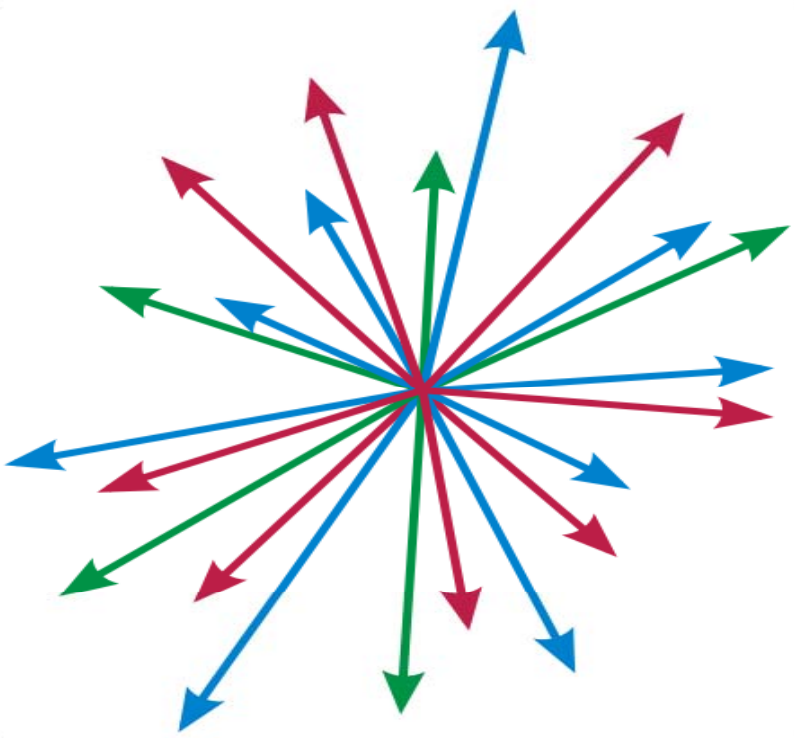
Usage > United States



Rank	Country	Pages/visit	Time on Site	New Visitors
1	California	5.85	0:04:44	0.7759
2	Florida	5.45	0:04:46	0.6901
3	Illinois	5.76	0:04:54	0.7568
4	New York	6.07	0:04:35	0.7461
5	Texas	6.68	0:05:05	0.794
6	Massachusetts	6.39	0:08:28	0.3498
7	New Jersey	5.03	0:03:19	0.8118
8	Pennsylvania	6.58	0:04:34	0.7885
9	Ohio	7.7	0:07:58	0.7443
10	Georgia	5.23	0:05:06	0.8754
11	Michigan	5.43	0:03:37	0.7955
12	Virginia	5.72	0:04:09	0.7598
13	Maryland	6.36	0:04:40	0.7015
14	Washington	4.65	0:03:44	0.843
15	Arizona	5.32	0:03:39	0.8232

Are transitional words really important?



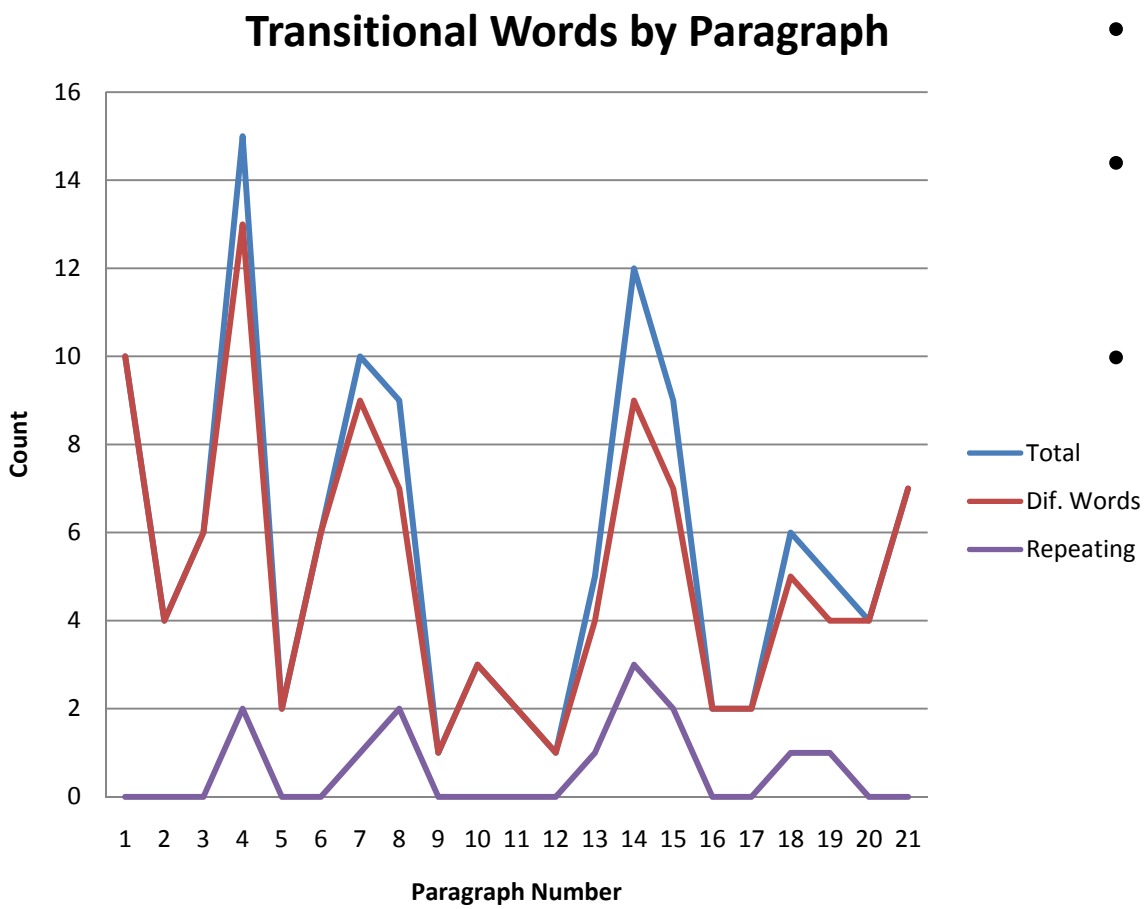


The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant

by

Andrew Gelman and Hal Stern

The American Statistician, November 2006, Vol. 60, No. 4



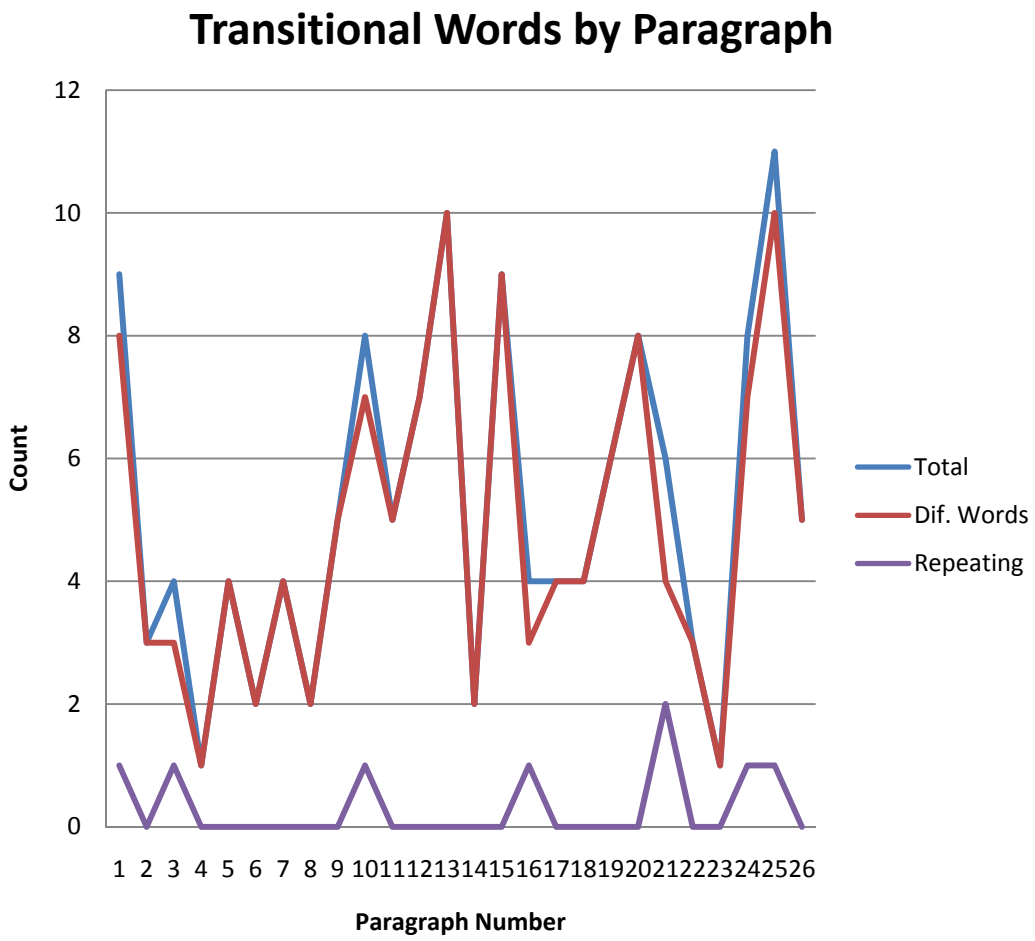
- 6 parts
- 21 paragraphs, 8 had repeating transitional words
- 121 transitional words, 108 account for non-repeating words within each paragraph
- Common repeating transitional words:
  - But, 5 paragraphs
  - As, 3 paragraphs
  - Between, 2 paragraphs
  - For example, 1 paragraph
  - Only , 1 paragraph
  - So forth, 1 paragraph

Technical Writing and Computer Programming

by

Judith Kaufman

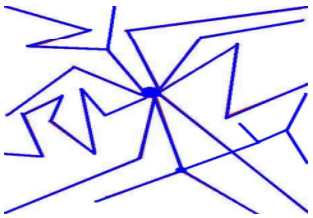
Transactions on Professional Communication, December 1988, Vol 31, No. 4.

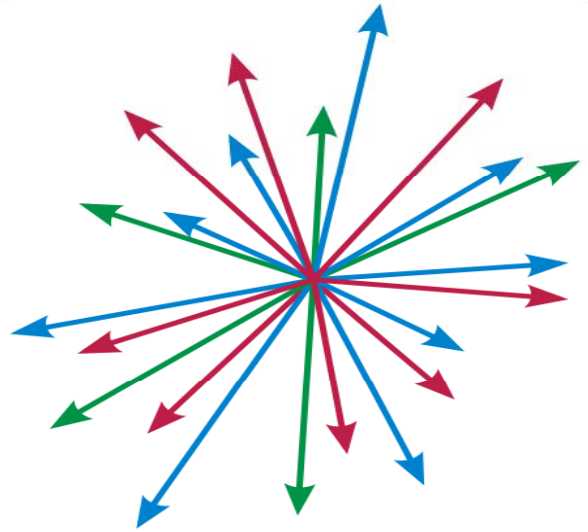


- 6 parts
- 26 paragraphs, 7 had repeating transitional words
- 135 transitional words, 127 account for non-repeating words within each paragraph
- Common repeating transitional words:
  - As, 2 paragraphs
  - Into, 2 paragraph
  - Between, 1 paragraph
  - Both, 1 paragraph
  - Just as, 1 paragraph
  - Thus, 1 paragraph

Forecasting Problem

- If transitional words are so important, shall we continue the development of ETWD?
  - ETWD has being used as an educational tool
  - Two journal articles show us the usage and relevance
- Decision:
  - Usage analysis
    - Web-metrics/web-analytics
    - Business analytics
    - Applied Statistics
  - Forecasting by ARMA models
    - Page views
    - Visits

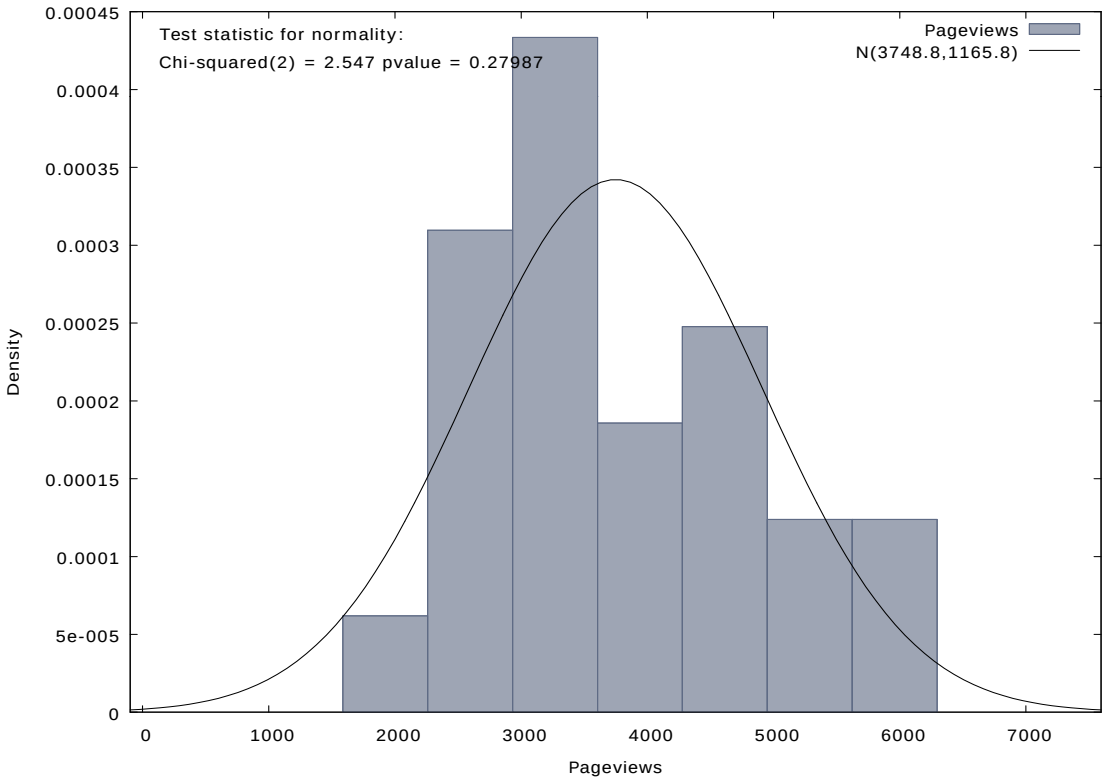




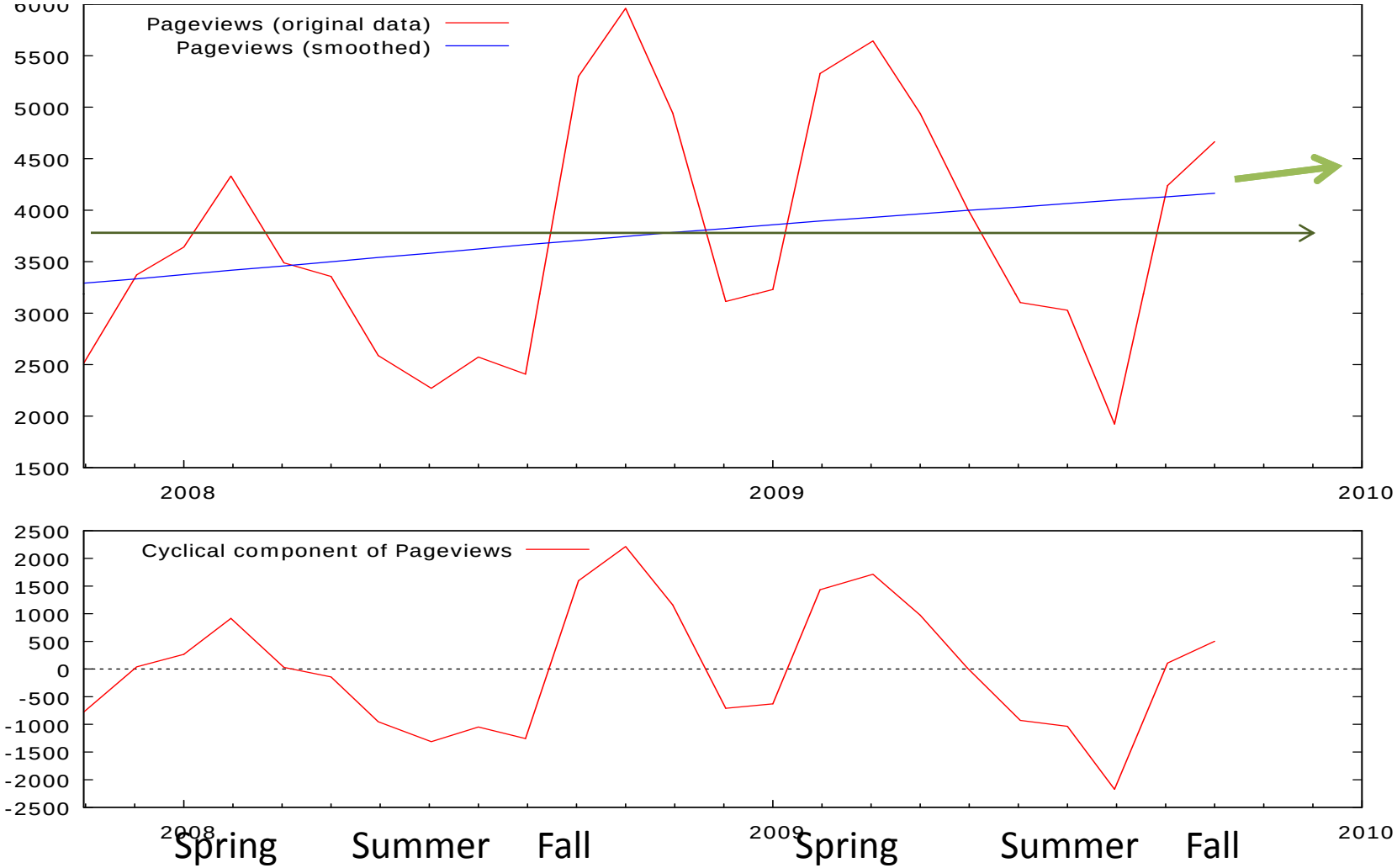
Page views

Summary statistics for the period 2007:11 - 2009:10

Mean	3748.8
Median	3430.0
Minimum	1923.0
Maximum	5960.0
Standard deviation	1165.8
C.V.	0.31099
Skewness	0.33030
Ex. kurtosis	-1.0452



Filtered Series for Page views: Hodrick & Prescott



14400

Unit Roots

Variable	Drift		Trend	
Levels \ Lags	Statistic	CV	Statistic	CV
1	-3.456	-1.729	-3.498	-3.24
2	-3.068	-1.74	-3.223	-3.24
3	-1.883	-1.753	-2.047	-3.24
First Diference				
1	-3.701	-1.734		
2	-4.443	-1.746		
3	-3.283	-1.761		

-ARIMA vs. ARMA
-Trends were not significant
-Process is stationary
-Selection criteria

Formula for the ARMA Model

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

Model Selection

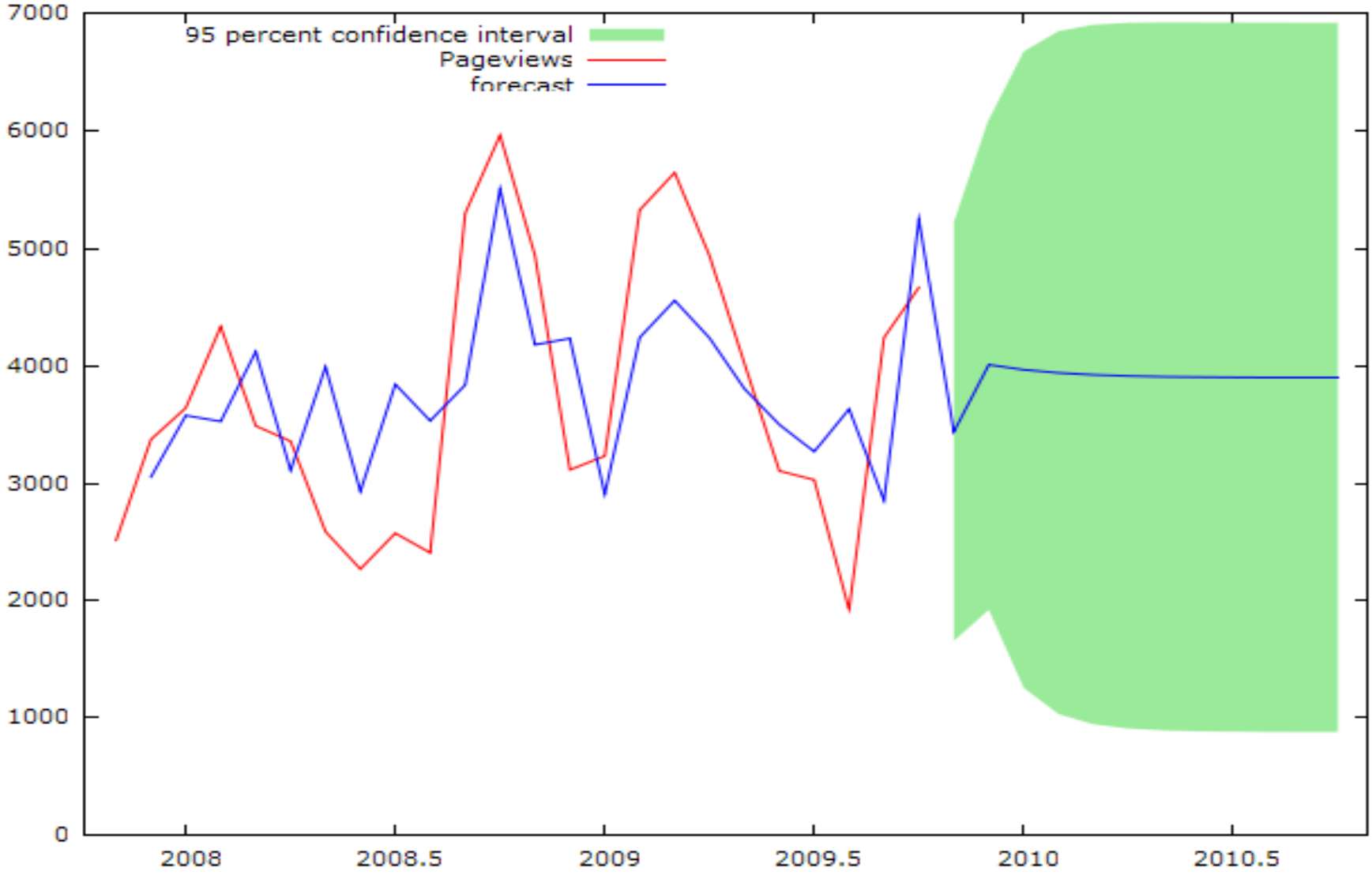
- (1)(2)
  - Log-likelihood -189.3678 Akaike criterion 386.7355
  - Schwarz criterion 391.2775 Hannan-Quinn 387.8778
  - All coefficients were significant
  - Predicted values above the mean
- (1)(3)
  - Log-likelihood -184.0443 Akaike criterion 376.0885
  - Schwarz criterion 380.6305 Hannan-Quinn 377.2308
  - All coefficients were significant but predicted values below the mean
  - Interval: reached negative values
- (1,2)(3)
  - Log-likelihood -176.4146 Akaike criterion 362.8292
  - Schwarz criterion 368.2844 Hannan-Quinn 364.1143
  - AR term 2 was not significant but interval reached negative values.
- (1,2)(2)
  - Log-likelihood -180.2015 Akaike criterion 370.4029
  - Schwarz criterion 375.8581 Hannan-Quinn 371.6880
  - AR term 2 and MA term 2 were not significant

Forecast of Monthly Page Views

Dependent variable: Page views
Estimates using the 23 observations 2007:12 - 2009:10

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	95% CONFIDENCE INTERVAL	
const	1516.2	646.93	2.3437	0.01909	248.239	2784.16
phi_1	0.610883	0.173347	3.524	0.00043	0.271129	0.950637
theta_2	-0.669686	0.218981	-3.0582	0.00223	-1.10	-0.24049

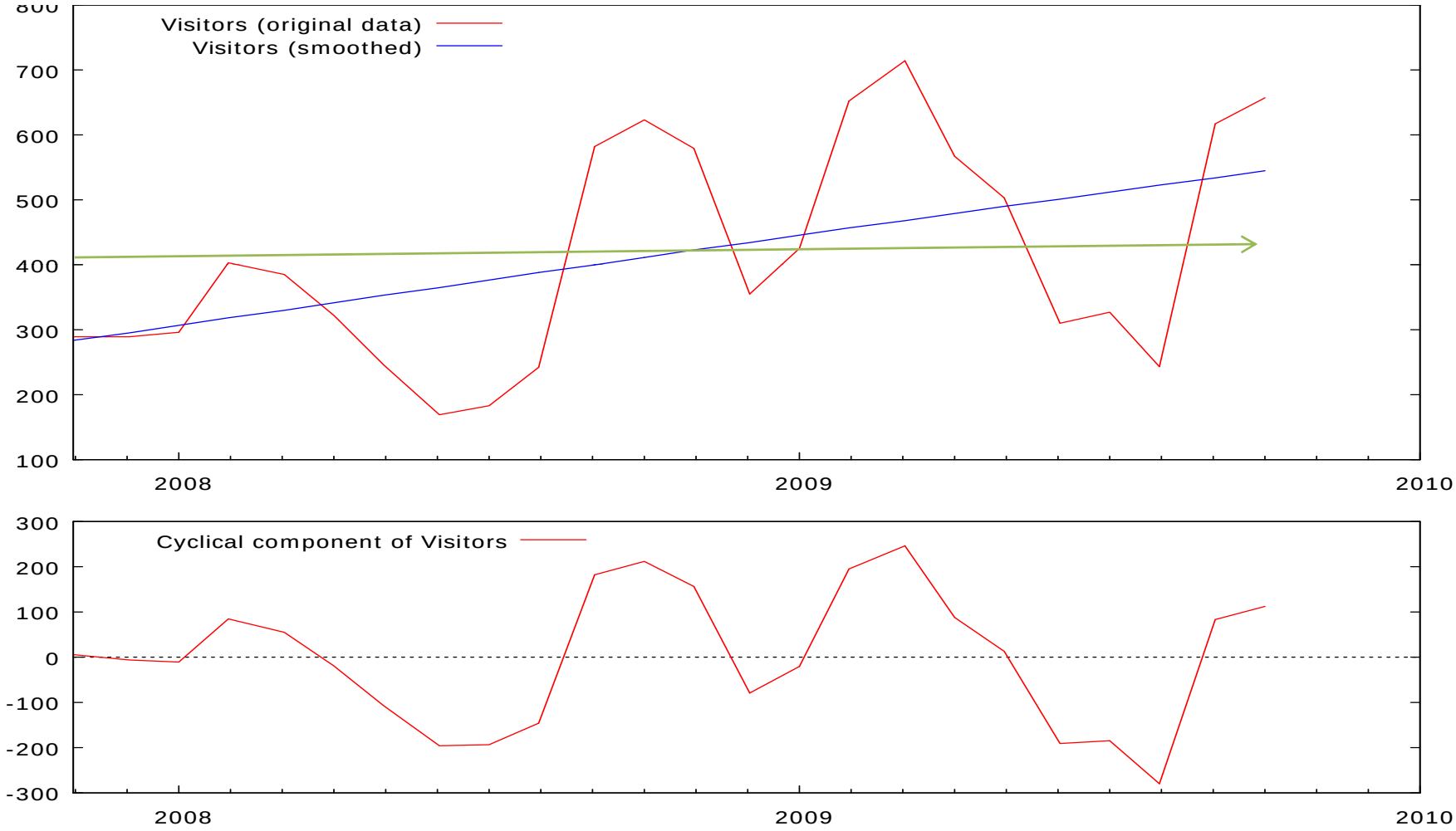
Forecast of Monthly Page Views



How well the model forecasts?

Month	Prediction	Std. error	95% Interval	Observed
2009:11	3436.94	910.909	1651.59 - 5222.29	5,387
2009:12	4007.47	1067.427	1915.35 - 6099.59	3,221
2010:01	3964.29	1383.475	1252.73 - 6675.86	3,853
2010:02	3937.92	1484.274	1028.80 - 6847.04	
2010:03	3921.81	1520.178	942.31 - 6901.30	
2010:04	3911.96	1533.361	906.63 - 6917.30	
2010:05	3905.95	1538.252	891.03 - 6920.87	

Visits


Visitors (original data)
Visitors (smoothed)
800
700
600
500
400
300
200
100
2008
2009
2010
Cyclical component of Visitors
300
200
100
0
-100
-200
-300
2008
2009
2010

Visits

Hodrick & Prescott Filter

Dependent variable: Visitors

	coefficient	std. error	t-ratio	p-value	
const	125.151	50.8205	2.463	0.0138	**
phi_1	0.536123	0.200636	2.672	0.0075	***
theta_3	-0.872983	0.208348	-4.190	2.79e-05	***
time	6.10663	3.49602	1.747	0.0807	*

Mean dependent var 421.2609 S.D. dependent var 170.2333
Mean of innovations -0.372145 S.D. of innovations 91.09357
Log-likelihood -136.4090 Akaike criterion 282.8180
Schwarz criterion 288.4955 Hannan-Quinn 284.2459

Forecast of Visits

Forecasted Visits

Date	Prediction	S.E.	Interval	Occurred
2009:11	698.91	91.094	520.37 - 877.45	766
2009:12	554.41	103.359	351.83 - 756.99	561
2010:01	634.10	106.624	425.12 - 843.08	640
2010:02	636.09	248.563	148.92 - 1123.27	
2010:03	643.27	276.178	101.97 - 1184.57	
2010:04	653.22	283.618	97.34 - 1209.10	
2010:05	664.66	285.720	104.66 - 1224.67	
2010:06	676.91	286.322	115.73 - 1238.09	
2010:07	689.58	286.495	128.06 - 1251.09	
2010:08	702.47	286.544	140.86 - 1264.09	
2010:09	715.50	286.559	153.85 - 1277.14	
2010:10	728.58	286.563	166.93 - 1290.24	

Conclusions

- ARMA models can be applied for forecasting the usage of a website, e.g. ETWD
- Cyclical components are in accordance with students' behavior and educational system in the USA, given that the majority of the user's base is originated from it.
- To continue or not to continue the development of ETWD based on the forecasted values of visits and page views?

  Yes!!!

Recommendations

- Content analysis
  - Descriptions
  - Definitions
  - Classification
  - Add examples to enhance learning
- Build a team
  - English
  - Linguistics
  - Programmer
- Apply for a grant!

Recommendations

- Classification
  - M.A. Student's thesis project
- Visual mapping
  - Part of a PhD dissertation
- Addition of Examples
  - Student workers (B.A.) from the English dept.
- Programming
  - Student workers (M.Sc.) from computer science dept.

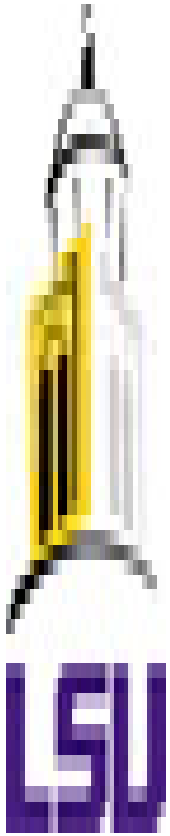
Forecasting Usability of the English Transitional Words Database

Carlos García

Seminar in the Department of Experimental Statistics

Louisiana State University A&M

February 17, 2010



Supporting Slides

Abstract Short

Proficiency in reading and writing depends on how well *transitional words* are interpreted; such interpretation determines message conveyance. Transitional devices link ideas, sentences and paragraphs providing coherence and unity to the text, they have an effect in the direction of the rhetorical objective by establishing relationships among ideas/objects. The English Transitional Words Database (ETWD) was created with the objective of organizing *transitional words* based on usage patterns rather than purely grammatical rules. Since its creation, the number of users has increased as well as the number of page views per month. It is necessary to determine whether or not ETWD should continue being developed. By using an ARMA model, it was concluded that it is very likely that usage of the ETWD will continue and thus more effort shall be invested on it.

Abstract Long

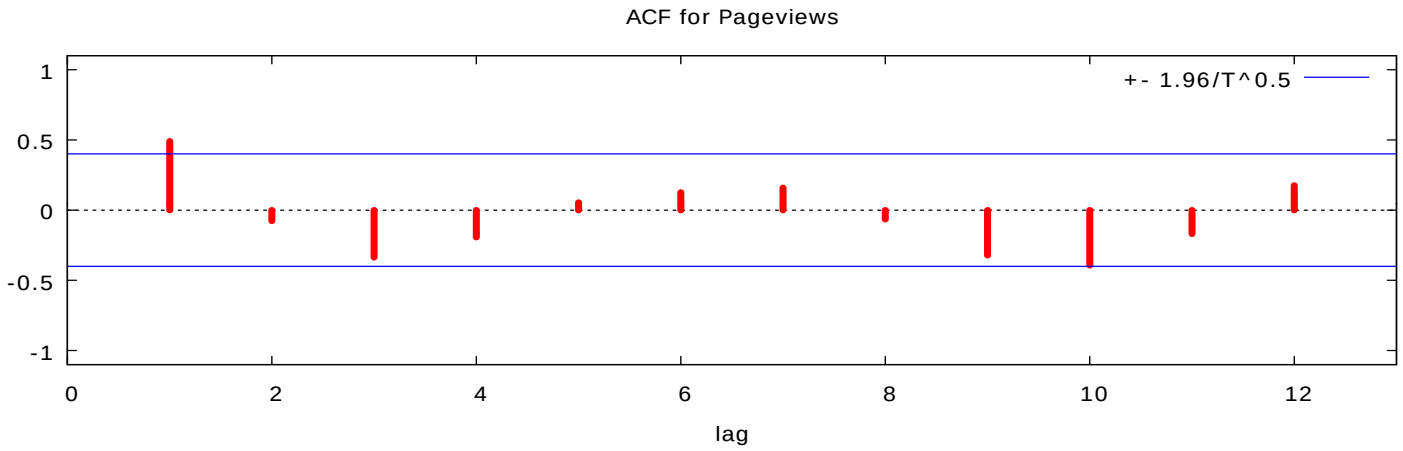
- Proficiency in reading and writing depends on how well transitional words are interpreted; such interpretation determines message conveyance. Transitional devices link ideas, sentences and paragraphs providing coherence and unity to the text, they have an effect in the direction of the rhetorical objective by establishing relationships among ideas/objects. The English Transitional Words Database (ETWD) was created with the objective of organizing transitional words based on usage patterns rather than purely grammatical rules.

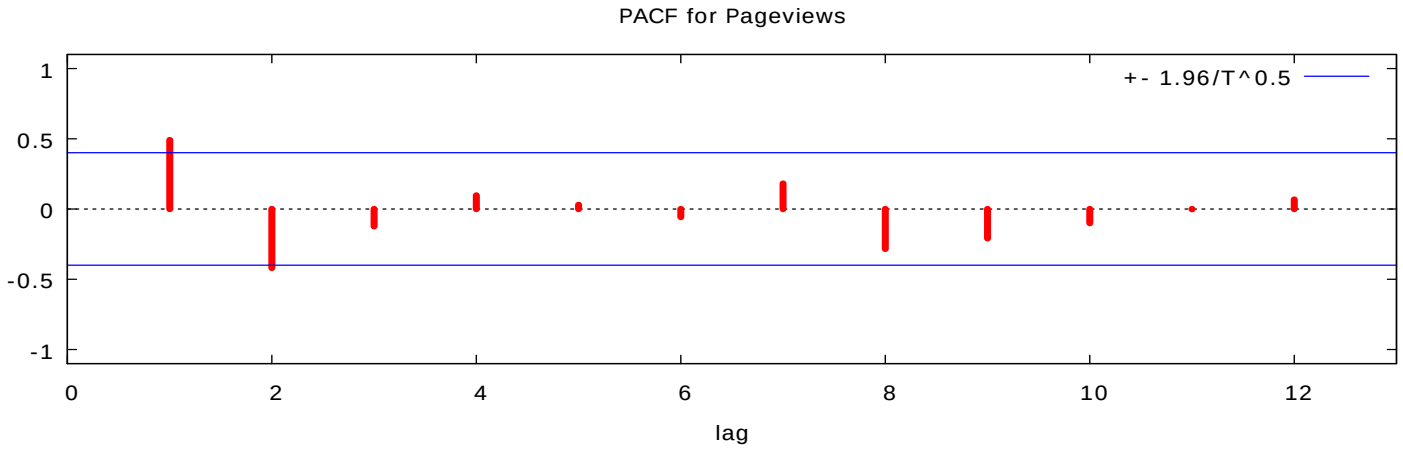
  ETWD is an educational tool that allows improvement of writing abilities by facilitating the choice of the best transitional word during the process of arguing for reaching the rhetorical goal. In addition, having more knowledge about transitional devices, readers are able to absorb more knowledge while the content is analyzed and the obtained information increases their well-being.

  Since the creation of ETWD, the number of users has increased as well as the number of page views per month; on average the website receives 3,450 monthly page views. Although the majority of the visitors originate from USA and Canada, developing countries visit the site as well, such as Philippines, South Korea, China, Malaysia, India, Russia and Latin America.

  By using an ARMA model, it was concluded that it is very likely that usage of the ETWD will continue and thus more effort shall be invested on it. By analyzing the content of the ETWD's website (http://www.carlosignacio.com/twd), it was determined that is imperative to continue its development by improving usability and accuracy of the description and classification of transitional words.

AC and PACF


ACF for Pageviews
+- 1.96/T^0.5
1
0.5
0
-0.5
-1
0
2
4
6
8
10
12
lag


PACF for Pageviews
+- 1.96/T^0.5
1
0.5
0
-0.5
-1
0
2
4
6
8
10
12
lag

Log-likelihood = -189.36777
Akaike information criterion (AIC) = 386.736
Schwarz Bayesian criterion (BIC) = 391.278
Hannan-Quinn criterion (HQC) = 387.878

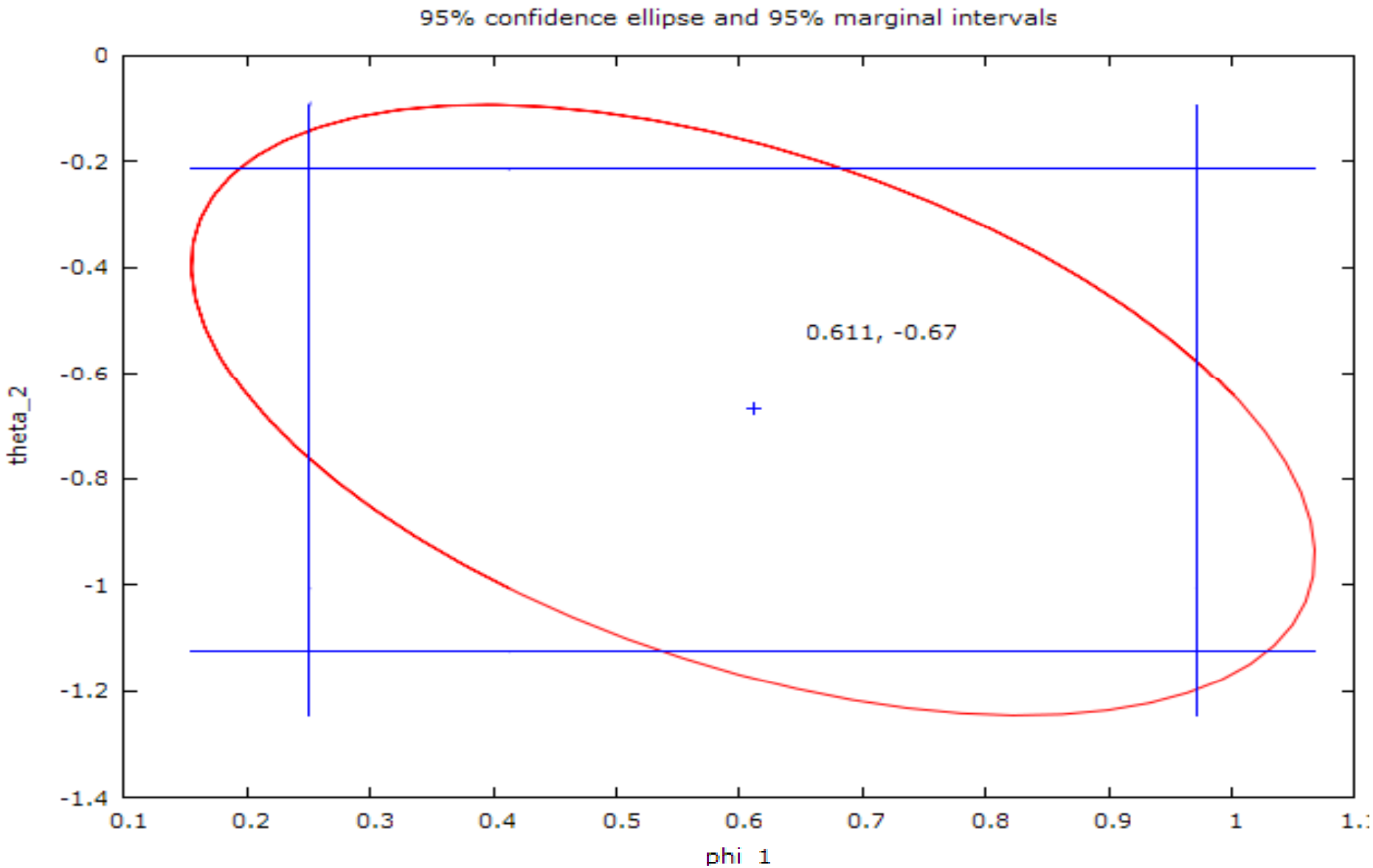
Test for normality of residual
Null hypothesis: error is normally distributed
Test statistic: Chi-square(2) = 1.34182
with p-value = 0.511242

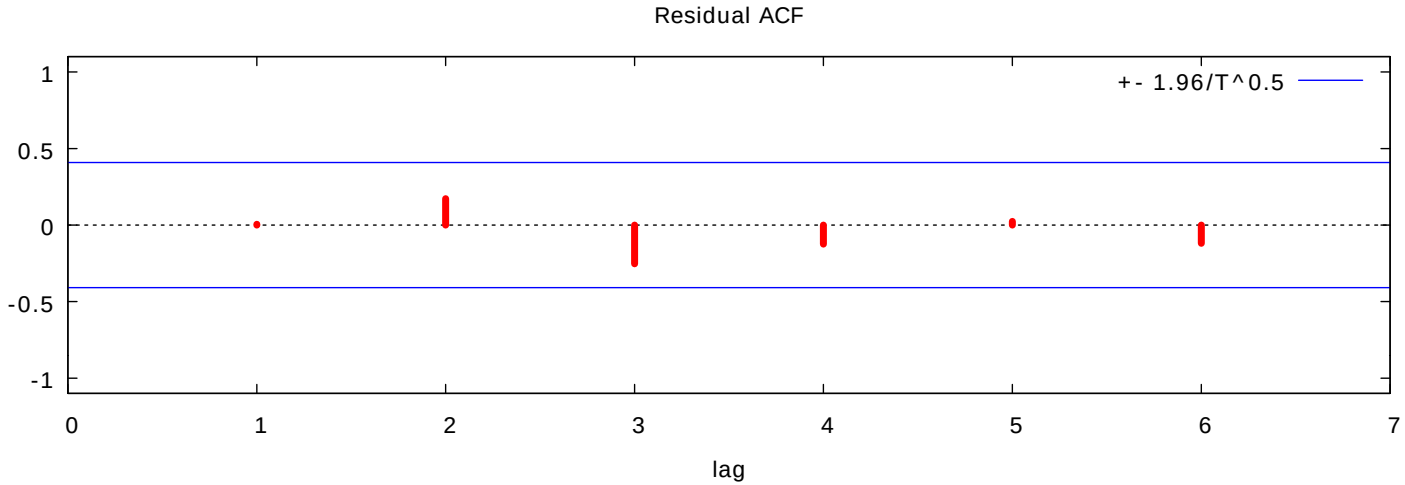
Test for ARCH of order 3

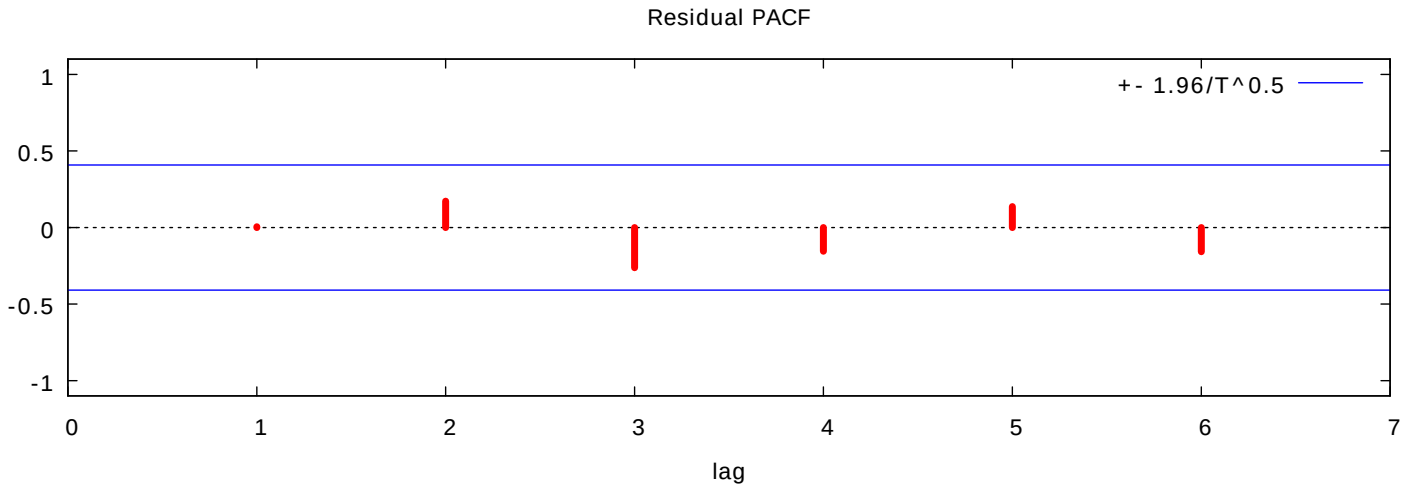
Null hypothesis: no ARCH effect is present

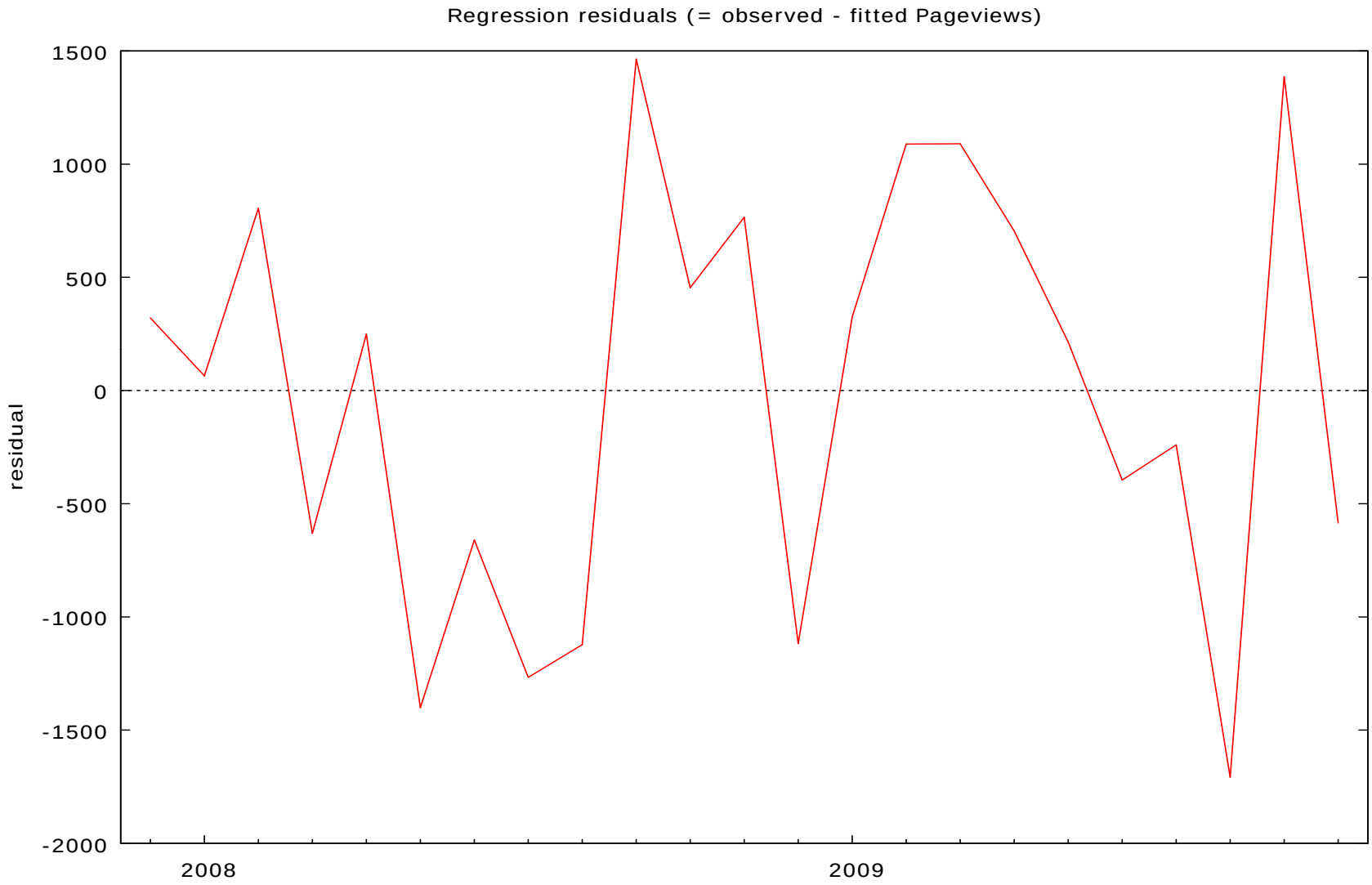
Test statistic: LM = 0.529754

with p-value = P(Chi-Square(3) > 0.529754) = 0.912308


95% confidence ellipse and 95% marginal intervals
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
theta_2
0.611, -0.67
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
phi_1


Residual ACF
+- 1.96/T^0.5
1
0.5
0
-0.5
-1
0
1
2
3
4
5
6
7
lag

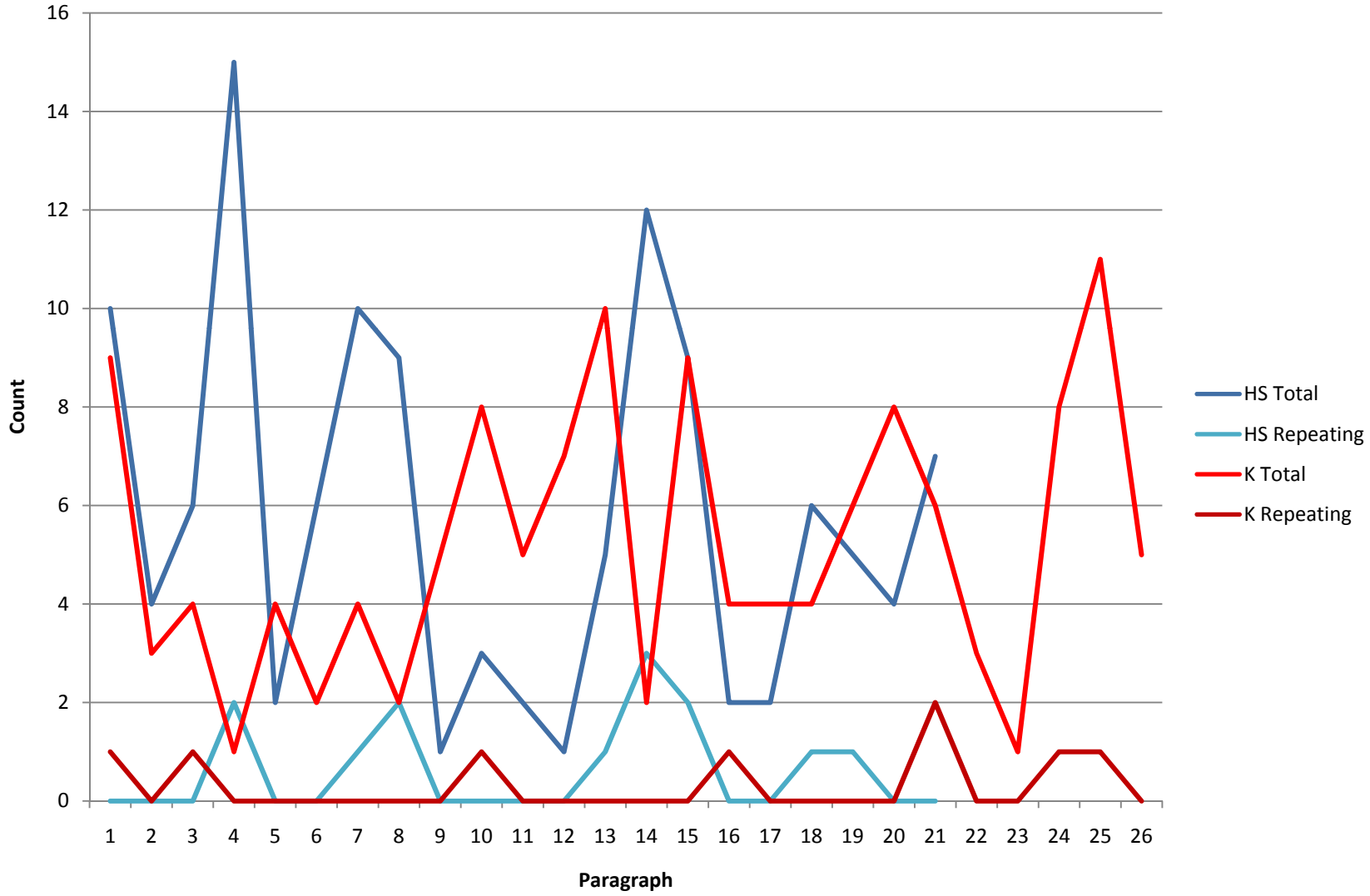

Residual PACF
+- 1.96/T^0.5
1
0.5
0
-0.5
-1
0
1
2
3
4
5
6
7
lag


Regression residuals (= observed - fitted Pageviews)
residual
1500
1000
500
0
-500
-1000
-1500
-2000
2008
2009

Formula for the Hodrick & Prescott Filter

$$MIN \sum_{t=1}^{T} (y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} [(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})]^2$$

Comparison of Papers: Gelman & Sterns and Kaufman



Model Selection

- (1)(2)
  - Log-likelihood -189.3678 Akaike criterion 386.7355
  - Schwarz criterion 391.2775 Hannan-Quinn 387.8778
  - All coefficients were significant
- (1)(3)
  - Log-likelihood -184.0443 Akaike criterion 376.0885
  - Schwarz criterion 380.6305 Hannan-Quinn 377.2308
  - All coefficients were significant
- (1,2)(3)
  - Log-likelihood -176.4146 Akaike criterion 362.8292
  - Schwarz criterion 368.2844 Hannan-Quinn 364.1143
  - AR term 2 was not significant
- (1,2)(2)
  - Log-likelihood -180.2015 Akaike criterion 370.4029
  - Schwarz criterion 375.8581 Hannan-Quinn 371.6880
  - AR term 2 and MA term 2 were not significant

(1)(3)

2009:11	5765.09	722.694	4348.64 - 7181.55
2009:12	3827.88	832.763	2195.70 - 5460.07
2010:01	3270.94	865.804	1573.99 - 4967.88
2010:02	3538.78	2209.177	-791.12 - 7868.69
2010:03	3692.14	2496.910	-1201.71 - 8585.99
2010:04	3779.94	2584.267	-1285.13 - 8845.01
2010:05	3830.21	2612.267	-1289.74 - 8950.16

(1,2)(3)

2009:11	5631.45	735.027	4190.82 - 7072.07
2009:12	3707.09	920.157	1903.61 - 5510.56
2010:01	2960.68	967.951	1063.53 - 4857.83
2010:02	3187.72	2282.029	-1284.98 - 7660.41
2010:03	3476.99	2789.331	-1990.00 - 8943.98
2010:04	3658.87	2924.988	-2074.00 - 9391.74
2010:05	3749.99	2953.428	-2038.62 - 9538.61
2010:06	3789.80	2958.224	-2008.21 - 9587.81
2010:07	3805.33	2958.874	-1993.95 - 9604.62
2010:08	3810.73	2958.942	-1988.69 - 9610.15
2010:09	3812.32	2958.947	-1987.10 - 9611.75
2010:10	3812.67	2958.947	-1986.76 - 9612.10

(1,2) (2)

2009:11	4533.07	873.089	2821.85 - 6244.29
2009:12	3797.39	1079.771	1681.08 - 5913.71
2010:01	3367.65	1152.462	1108.86 - 5626.43
2010:02	3481.48	1154.915	1217.88 - 5745.07
2010:03	3813.47	1190.355	1480.42 - 6146.52
2010:04	3989.05	1201.896	1633.38 - 6344.73
2010:05	3924.33	1202.785	1566.91 - 6281.74
2010:06	3775.43	1209.789	1404.29 - 6146.57
2010:07	3704.61	1211.686	1329.75 - 6079.47
2010:08	3739.41	1211.966	1364.00 - 6114.82
2010:09	3805.79	1213.365	1427.64 - 6183.94
2010:10	3833.92	1213.672	1455.16 - 6212.67

Acces

- Last 6 months
  - 24.72% Direct Traffic
  - 45.05% Referring Sites
  - 30.22% Search Engines
- 2010
  - 41.72% Direct Traffic
  - 37.75% Referring Sites
  - 20.52% Search Engines

Model 1: OLS estimates using the 24 observations 2007:11-2009:10
Dependent variable: **Visitors**

	coefficient	std. error	t-ratio	p-value
const	335.750	43.5503	7.709	1.08e-07 ***
dyear	160.000	61.5895	2.598	0.0164 **

Mean dependent var 415.7500 S.D. dependent var 168.6662
Sum squared resid 500710.5 S.E. of regression 150.8627
R-squared 0.234751 Adjusted R-squared 0.199967
F(1, 22) 6.748810 P-value(F) 0.016424
Log-likelihood -153.4033 Akaike criterion 310.8066
Schwarz criterion 313.1627 Hannan-Quinn 311.4316
rho 0.485497 Durbin-Watson 1.023134

Model 2: OLS estimates using the 24 observations 2007:11-2009:10
Dependent variable: **Pageviews**

	coefficient	std. error	t-ratio	p-value
const	3483.50	334.689	10.41	5.80e-010 ***
dyear	530.500	473.322	1.121	0.2745

Mean dependent var 3748.750 S.D. dependent var 1165.837
Sum squared resid 29572471 S.E. of regression 1159.398
R-squared 0.054016 Adjusted R-squared 0.011016
F(1, 22) 1.256195 P-value(F) 0.274459
Log-likelihood -202.3461 Akaike criterion 408.6923
Schwarz criterion 411.0484 Hannan-Quinn 409.3173
rho 0.462303 Durbin-Watson 1.042468

Visits

Descriptive Statistics

Summary Statistics, using the observations 2007:11 - 2010:10
for the variable 'Visitors' (24 valid observations)

Mean	415.75
Median	370.00
Minimum	169.00
Maximum	714.00
Standard deviation	168.67
C.V.	0.40569
Skewness	0.29810
Ex. kurtosis	-1.3053

First Differences