Μέθοδοι και Τεχνικές Ανάλυσης Χρονολογικών Σειρών

1. Εισαγωγή

Στην συγκεκριμένη εργασία θα ασχοληθούμε με την εξαγωγή των δεδομένων τα όποια βρίσκονται σε ένα xml file, το όποιο παρέχετε από το http://dblp.uni-trier.de όπου είναι ένα computer science bibliography. Τα δεδομένα που θέλουμε να εξάγουμε από το συγκεκριμένο data set περιέχει το πλήθος των άρθρων σε περιοδικά , διατριβές ,δημοσιεύσεις κ.α. που έχουν γίνει στο κλάδο του computer science ανά έτος. Στην συνεχεία θα ασχοληθούμε με τη ανάλυση διαφορών μοντέλων χρονολογικών σειρών όπου ως χρονολογικές σειρές ορίζεται ένα σύνολο παρατηρήσεων μιας μεταβλητής οι οποίες έχουν ληφθεί σε ίσα χρονικά διαστήματα .Για τα συγκεκριμένα δεδομένα χρησιμοποιήθηκαν μοντέλα χρονολογικών σειρών όπως την στοχαστική προσέγγιση ARIMA ,εκθετική εξομάλυνση καθώς και το υπόδειγμα τυχαίας διαδρομής . Δηλαδή μοντέλα τα όποια έχουν υψηλό accuracy ,καθώς και να έχουν την ικανότητα για σωστή πρόβλεψη του μέλλοντος. Οι χρονολογικές σειρές αναλύονται σύμφωνα με την Τάση, Εποχικότητα, Κυκλικότητα, Τυχαιότητα και αναπτύσσεται η πρόβλεψη από τη σύνθεση αυτών των συνιστωσών. Γενικά οι μέθοδοι προβλέψεις μπορούν διαχωριστούν σε ποιοτικές και ποσοτικές. Οι ποιοτικές προβλέψεις στηρίζονται σε ποιοτικά δεδομένα , όπου οι επιστήμονες αναλυτές χρησιμοποιούν την εμπειρία ή την κρίση τους αλλά και στατιστικές μεθόδους (DELPHI) ενώ οι ποσοτικές προβλέψεις στηρίζονται μόνο σε στατιστικές μεθόδους και αφορούν ποσοτικές μεταβλητές. Στην συγκεκριμένη εργασία όπως είδαμε και παραπάνω θα ασχοληθούμε με ποσοτικές προβλέψεις

2. Ανάλυση Χρονολογικών Σειρών

Στην ανάλυση ποσοτικών χρονολογικών σειρών έχουμε δυο ειδή μοντέλων, το προσθετικό και το πολλαπλασιαστικό μοντέλο, όπου το προσθετικό είναι της μορφής :

$$Y_t = T_t + S_t + C_t + R_t$$

Καθώς και το πολλαπλασιαστικό :

$$Y_t = T_t * S_t * C_t * R_t$$

Όπου Y_t είναι η τιμή της μεταβλητής, S_t είναι η εποχιακή συνιστώσα , C_t η κυκλική συνιστώσα και R_t είναι η τυχαία συνιστώσα .Επίσης μια σημαντική έννοια για την ανάλυση χρονολογικών σειρών είναι η στασιμότητα. Μια στοχαστική διαδικασία χαρακτηρίζεται ως στάσιμη όταν οι στατιστικές της ιδιότητες δεν επηρεάζονται από μια μεταβολή στην αρχή του χρόνου . Δηλαδή , οι στατιστικές ιδιότητες των N παρατηρήσεων με αρχή το t $(y_t, y_{t+1}, \dots, y_{t+N-1})$ είναι οι ίδιες με τις στατιστικές ιδιότητες των N παρατηρήσεων με αρχή την περίοδο t+k $(y_{t+k}, y_{t+k+1}, \dots, y_{t+k+N-1})$. Γενικά μια χρονολογική σειρά θα είναι στάσιμη αν ο μέσος και η διακύμανση της δεν μεταβάλλονται με το χρόνο και η συνδιακύμανση μεταξύ των τιμών της σε δυο χρονικά σημεία και όχι από τον ίδιο το χρόνο. Αν μια χρονολογική σειρά είναι στάσιμη, τότε για όλα τα t θα ισχύουν :

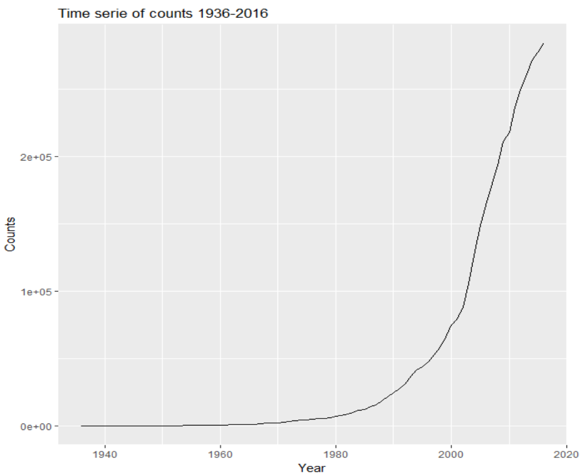
- $E(y_t) = \mu_y$,
- $var(y_t) = E[y_t - E(y_t)]^2$
- $cov(y_t, y_{t+k}) = cov(y_{t+m}, y_{t+m+k})$

όπου οι δυο πρώτες συνθήκες υποδηλώνουν σταθερό μέση και σαθρή διακύμανση . Η τρίτη δηλώνει ότι η συνδιακύμανση μεταξύ δυο οποιονδήποτε τιμών της y_t που απέχουν κ περιόδους (αυτοσυνδιακυμαση) είναι συνάρτηση μόνο του k ,δηλαδή της χρονικής υστέρησης ή προήγησης των δυο αυτών τιμών .

3. Επεξεργασία Δεδομένων

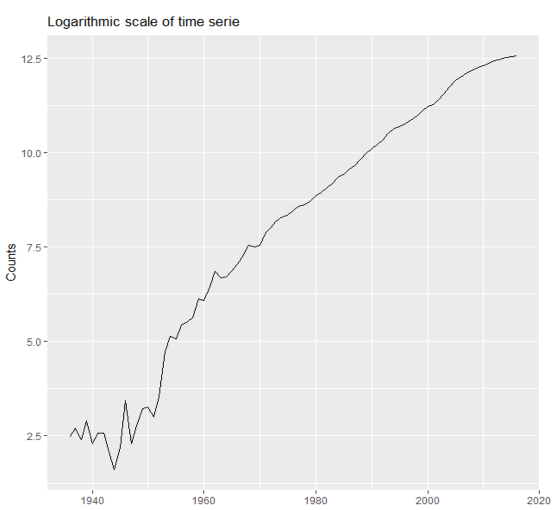
Τα δεδομένα ήταν κοντά στα 2GB πράγμα που σημαίνει πως μια γλωσσά προγραμματισμού όπως R, Python κ.α. είναι δύσκολο και όχι τόσο καλή λύση να ανοίξουν το συγκεκριμένο αρχείο. Τα δεδομένα ανοιχτήκαν από το command line ενός Ubuntu Server καθώς και η επεξεργασία των δεδομένων που θα χρησιμοποιήσουμε ,διότι στο xml αρχείο που είχαμε, περιείχε περίπου 55 εκατομμύρια γραμμές εκ των οποίων εμάς μας ενδιέφεραν μόνο οι 4.092,332 γραμμές όπου είχαν την χρονολογία. Αφού κάναμε εξαγωγή ένα νέο αρχείο με της γραμμές που περιέχουν την χρονολογία έπειτα συνεχίσαμε διαβάζοντας το αρχείο αυτό στην R και υπολογίσαμε τις συχνότητες εμφάνισης κάθε χρονιάς .Έπειτα συνεχίσαμε με την ανάλυση της χρονολογικής σειράς καθώς και την πρόβλεψη της.

Διάγραμμα 3.1



Στο παραπάνω διάγραμμα βλέπουμε τα δεδομένα μας και παρατηρούμε ότι υπάρχει μια αυξητική τάση. Έχουμε βγάλει εκτός μερικές χρονιές (2017 έως 2019) για τον λόγο ότι υπήρχε μια καθοδική τάση, πράγμα το όποιο έχουμε υπόθεση ότι είναι εσφαλμένο και δεν αντανακλά την αληθινή εικόνα . Αυτό μπορεί να οφείλεται στο ότι δεν έχουν περαστεί ακόμα όλα τα δεδομένα για αυτά τα χρόνια .Κάτι που δεν μπούμε να δούμε στο παραπάνω διάγραμμα είναι μετατοπίσεις το όποιο οφείλεται στην κλίμακα που έχουμε απεικονίσει . Μια καλή κλίμακα για να φάνουνε οι μετατοπίσεις αυτές, είναι η λογαριθμική κλίμακα .

Διάγραμμα 3.2



Στο παραπάνω διάγραμμα παρατηρούμε αρκετές μετατοπίσεις από το 1936 έως 1970 περίπου τα όποια πριν φαινόντουσαν να έχουν μια σταθερή αυξητική τάση.

4. ARIMA

Αυτοπαλινδρομο Υπόδειγμα

Αυτοπαλινδρομο υπόδειγμα p τάξης συμβολίζετε με AR(p) και εκφράζετε από την σχέση :

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t$$

Όπου α_ι *με* $\mathrm{i} \in [0, \mathrm{p}]$ *είναι η σταθεροί παράμετροι και* ε_t *ο όποιος μετρά τυχαία σφάλματα. Στο αυτοπαλινδονουμενο υπόδειγμα , η εξαρτώμενη μεταβλητή* Y_t *παλινδρομείται στης προηγούμενες τιμές της . Η τάξη του αυτοπαλινδρονουμενου υποδείγματος συμβολίζεται με p και προσδιορίζει το μήκος της υστέρησης .*

Υπόδειγμα Κινητού Μέσου

Ένα υπόδειγμα κινητού μέσου q τάξης συμβολίζεται με MA(q) και εκφράζετε από την σχέση :

$$Y_\tau = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

Όπου θ_i *με* $\mathrm{i} \in [0, q]$ *οι σταθεροί παράμετροι και* ε_t *τα τυχαία σφάλματα .Στην διαδικασία κινητού μέσου η χρονολογική σειρά* Y_t *θεωρείται ότι δημιουργείται ως ένας σταθμικός μέσος τυχαίων σφαλμάτων των q προηγουμένων περιόδων .*

Ένα αυτοπαλινδρομο ολοκληρωμένο υπόδειγμα κινητού μέσου ARIMA(p,d,q) προκύπτει από τον συνδυασμό των αυτοπαλινδρομον διαδικασιών AR(p) και των διαδικασιών κινητού μέσου MA(q), όπου d είναι ο αριθμός των διαφόρων που απαιτούνται προκείμενου να μετατραπεί η σειρά σε στάσιμη.

Έλεγχος Στασιμότητας

Κάνοντας τον έλεγχο στασιμότητας Dickey-Fuller tests και Phillips-Perron Unit Root Test στα *δεδομένα μας. Όπου* Το κριτήριο των Dickey-Fuller βασίζεται στον ακόλουθο έλεγχο:

H_0:α=1 Η H_0 γίνεται δεκτή αν η t στατιστική *συνάρτηση* του συντελεστή α είναι μικρότερη από την t στατιστική *συνάρτηση* των Dickey-Fuller (αντίστοιχα για το Phillips-Perron).αν δεν έχουμε ένδειξης να απορριφτεί η H_0 τότε η σειρά δεν είναι στάσιμη

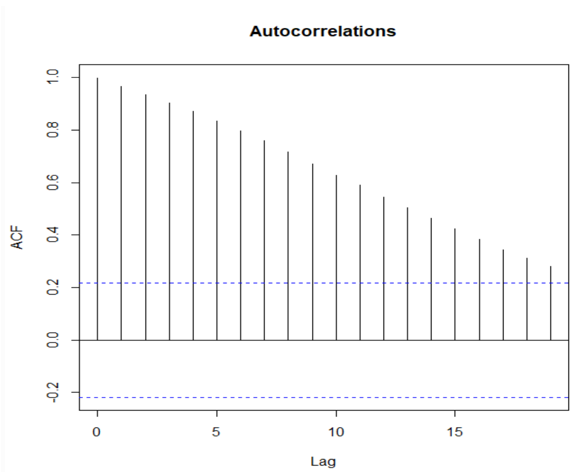
H_1:α<1 Τότε συμπεραίνουμε ότι έχουμε μοναδιαία ρίζα και άρα η σειρά είναι στάσιμη

Πινάκας 4.1

Έλεγχος	***P-value***	***Στατιστική συνάρτηση-T***
Dickey-Fuller	*0.8379*	*-1.3596*
Phillips-Perron Unit Root	*0.737*	*-6.4766*

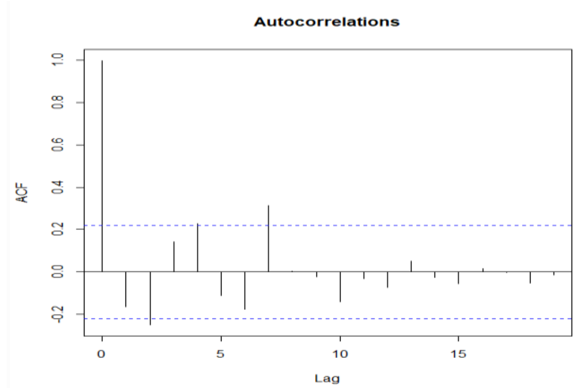
Και τα δύο τεστ συμφωνούν στο γεγονός μη επαρκούς ένδειξης απόρριψης της μηδενικής υπόθεσης, η χρονοσειρά ενδέχεται να έχει μοναδιαία ρίζα. Αυτό προκύπτει διότι το p-value είναι μεγαλύτερο από το επίπεδο σημαντικότητας 5% που ορίζουμε εμείς ότι θέλουμε να έχουμε ως σφάλμα. Από τους δυο αυτούς ελέγχους συμπεραίνουμε ότι η χρονοσειρα μας δεν είναι στάσιμη όποτε δεν πληρούνται οι προϋποθέσεις για το μοντέλο ARIMA.Περά από τους δυο αυτούς τους ελέγχους μπορούμε να διαπιστώσουμε και από την συνάρτηση αυτοσυσχέτισης (correlogram) που δίνετε στο διάγραμμα 4.1. Οι αυτοσυσχετισεις 'AC' φθίνουν με πολύ αργό ρυθμό που δηλώνει μη στασιμότητα. . Μια τεχνική για να γίνει η χρονολογική σειρά στάσιμη είναι το να πάρουμε τις πρώτες διαφορές της σειράς , δηλαδή ορίζουμε ως πρώτες διαφορές $y'_t = y_t - y_{t-1}$.

Διάγραμμα 4.1

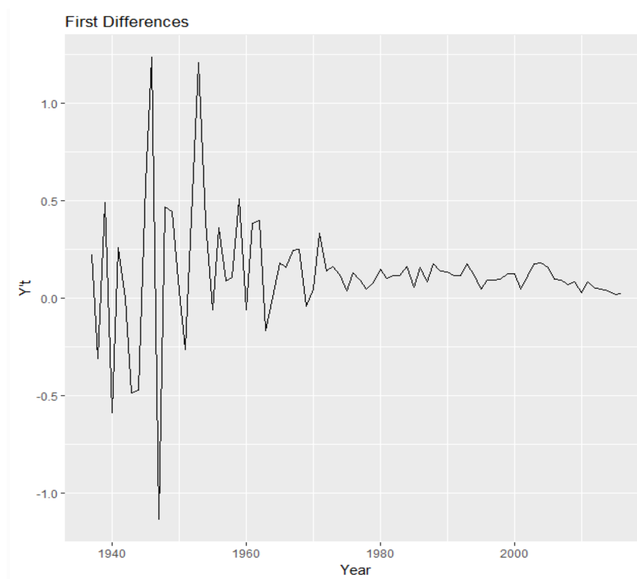


Αφού υπολογίσουμε τις διαφορές αυτές και ξανά κάνουμε τους δυο ελέγχους τότε παρατηρούμε πως δεν θα δεχτούμε τις μηδενικές υπόθεσης διότι τα p-value είναι 0.04731 και μικρότερα από 0.01 αντίστοιχα για τους δυο ελέγχους και οι στατιστικές συναρτήσεις είναι -3.502, -82.097 ,κάτι που φαίνετε και από το διάγραμμα αυτοσυσχέτισης παρακάτω. Επόμενος αφού έχουμε πλέον μια στάσιμη χρονολογική σειρά μπορούμε να συνεχίσουμε με την εκτίμηση των παραμέτρων του μοντέλου ARIMA(p,1,q).

Διάγραμμα 4.2



Διάγραμμα 4.3 Η χρονοσειρά με πρώτες διαφορές y_t .



Μέθοδος Box-Jenkins

Η μέθοδος Box-Jenkins περιγράφεται από τρία βασικά σταδία :

1. *Ταυτοποίηση*
2. *Εκτίμηση*
3. *Διαγνωστικός Έλενος*

Στο στάδιο της ταυτοποίησης γίνεται ο καθαρισμός των τιμών p,d,q του μοντέλου ARIMA με βάση της πληροφορίας που παρέχει το δείγμα. Στην αρχή καθορίζεται ο αριθμός των διαφόρων d που απαιτούνται ώστε να γίνει η σειρά στάσιμη .Προφανώς η εκτίμηση για την τιμή του d είναι μηδέν διότι τι δουλεία αυτή την κάναμε εμείς προηγούμενος . Για τον έλεγχο στασιμότητας εξετάζεται η δειγματική συνάρτηση αυτοσυσχέτισης. Μετά το στάδιο της *ταυτοποίησης έπεται το στάδιο της εκτίμησης όπου εκεί εκτιμούνται το p,d,q όπου έχουμε αναφέρει .Αφού ολοκληρωθεί το στάδιο της εκτίμησης συνεχίζει στο στάδιο του διαγνωστικού έλεγχου του οποίου ο σκοπός είναι να χρησιμοποιηθεί το μοντέλο για μελλοντικές πρόβλεψεις .Αρχικά διαχωρίσουμε τα δεδομένα μας σε δεδομένα εκπαίδευσης και δεδομένα έλεγχου, ,με δεδομένα εκπαίδευσης να είναι τις χρονολογίες [1938,2007] και τα δεδομένα έλεγχου [2008,2015] . Παρατηρούμε πως τα δεδομένα μας αρχικά είχαν ως αφετηρία το 1936 και ήταν μέχρι και το 2016 αλλά λόγο των δευτέρων διαφορών που πήραμε μερικές χρονιές εξαφανιστήκαν .*

Από την μέθοδο Box –Jenkins πήραμε ως εκτίμηση τον (p,d,q)=(5,1,3)

Με συντελεστές του μοντέλου τους βλέπουμε στον παρακάτω πινάκα .

Πινάκας 4.2

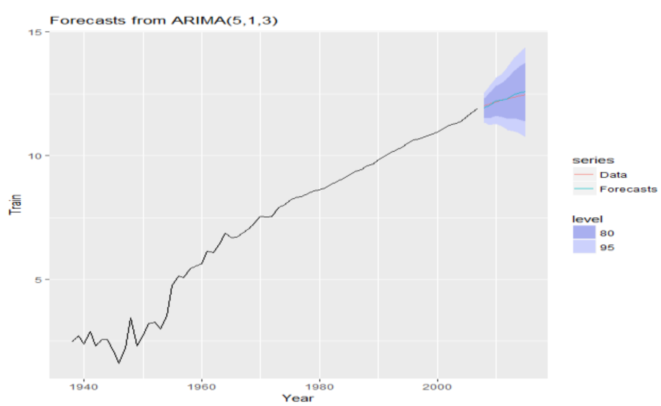
```
Series: Train
ARIMA(5,1,3)

Coefficients:
          ar1      ar2     ar3     ar4     ar5     ma1     ma2      ma3
      -0.3207  -0.3206  0.8130  0.3829  0.2869  0.2074  0.2034  -0.7948
s.e.   0.1792   0.1374  0.0842  0.1207  0.1415  0.1598  0.1405   0.1551

sigma^2 estimated as 0.09229:  log likelihood=-13.71
AIC=45.42   AICc=48.47   BIC=65.53
```

Στο παρακάτω διάγραμμα βλέπουμε την χρονολογική σειρά μαζί με την πρόβλεψη όπου έγινε ,καθώς και τα πραγματικά δεδομένα για τις χρονιές της πρόβλεψης . Με μπλε χρώμα βλέπουμε το διάστημα εμπιστοσύνης που μας δίνετε για την πρόβλεψη.

Διάγραμμα. 4.2



Στον πινάκα 4.3 βλέπουμε τα αριθμητικά αποτελέσματα που πήραμε από την πρόβλεψη μας ,τις πραγματικές τιμές καθώς και δυο διάστημα εμπιστοσύνης, το ένα για 95% και ένα για 80% .

***Πίνακας 4.3** Αποτελέσματα εκτίμησης 2008-2015*

Point	Forecast	Test	Lo 80	Hi 80	Lo 95	Hi 95
2008	11.93218	12.00633	11.53812	12.32624	11.32952	12.53484
2009	12.02977	12.10086	11.50630	12.55325	11.22919	12.83036
2010	12.20876	12.17299	11.60106	12.81645	11.27937	13.13815
2011	12.25172	12.25917	11.55008	12.95336	11.17866	13.32478
2012	12.31413	12.29070	11.47545	13.15281	11.03148	13.59678
2013	12.46969	12.37478	11.50534	13.43403	10.99485	13.94452
2014	12.53125	12.42728	11.47041	13.59210	10.90883	14.15368
2015	12.58018	12.47365	11.38897	13.77140	10.75838	14.40199

Παρατηρούμε πως η πρόβλεψη του μοντέλου είναι αρκετά ακριβής σε σύγκριση με τα πραγματικά δεδομένα . Η αξιοπιστία μετράται με μια αντικειμενική συνάρτηση (πχ RMSE root mean squared error), απόκλιση της πρόβλεψης από την πραγματική τιμή στην περίοδο πρόβλεψης

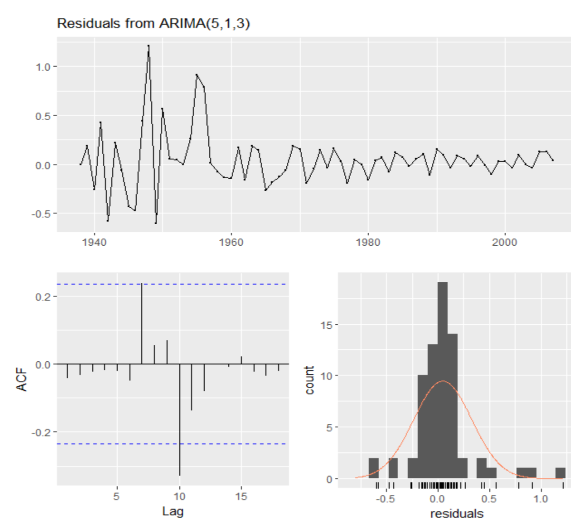
Πίνακας 4.4

	ME	RMSE	MAE	MPE	MAPE	MASE
Training set	0.04590220	0.28358884	0.17729671	0.5523769	4.6898068	0.7382742
Test set	-0.02649055	0.07382552	0.06466104	-0.2095381	0.5259818	0.2692525

	ACF1	Theil's U
Training set	-0.04218276	NA
Test set	0.52641365	1.039054

Από ότι βλέπουμε η τιμή του RMSE είναι αρκετά μικρή και για το Train set αλλά και για το test set όποτε η ακρίβεια μας είναι αρκετά μεγάλη για τον λόγο ότι η απόκλιση από της πραγματικές τιμές είναι αρκετά μικρή ποσότητα.

Διάγραμμα 4.3



Από ότι παρατηρούμε από το παραπάνω διάγραμμα τα κατάλοιπα της παλινδρόμησης ακολούθου περίπου κανονική κατανομή, για την ακρίβεια έχει πιο «παχιές ουρές», πράγμα αναμενόμενο για μια χρονολογική σειρά .

5. Εκθετική Εξομάλυνση

Οι μέθοδοι εκθετικής εξομάλυνσης βασίζονται στην εκθετική μείωση της βαρύτητας που δίνεται στα στοιχεία των προηγούμενων περιόδων. Αυτές οι μέθοδοι συνήθως χρησιμοποιούνται σε περιπτώσεις όπου ο χρονικός που θα ορίσουμε για την πρόβλεψη είναι σχετικά μικρός ενώ δεν υπάρχουν διαθέσιμες πληροφορίες για την αιτιοκρατική σχέση που συνδέει την προς πρόβλεψη μεταβλητή και τους ανεξάρτητους παράγοντες που την επηρεάζουν. Επίσης σημαντικό είναι ότι η εκθετική εξομάλυνση χαρακτηρίζονται από την εξομάλυνση των τυχαίων διακυμάνσεων που μπορεί να παρουσιάζουν τα διάφορα στοιχεία των χρονοσειρών (οριζόντιο, τάσης, εποχικό και κυκλικό).

D_t είναι η πραγματική ζήτηση την περίοδο t

F_{t+1} είναι η πρόβλεψη για την επόμενη περίοδο

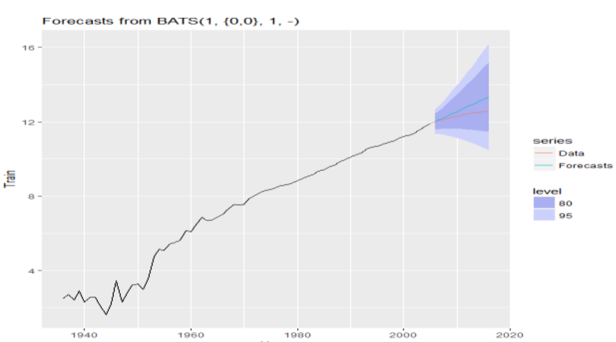
α=Σταθερά εξομάλυνσης (μεταξύ 0 και 1), συνήθως μεταξύ 0,01 και 0,3

$$F_{t+1} = F_t + \frac{D_t - F_t}{N} = \frac{1}{N}D_t + (1 - \frac{1}{N})F_t$$

$$F_t = aD_t + (1 - a)F_t = F_t + a(D_t - F_t)$$

Στην συγκεκριμένο πρόβλημα χρησιμοποίησα το μοντέλο Exponential Smoothing State Space Model With Box-Cox Transformation, ARMA Errors, Trend And Seasonal Components.To Box-Cox είναι μια οικογένεια μετασχηματισμών μέσα στους οποίους μπορούμε να προσδιορίσουμε τον πλέον κατάλληλο μετασχηματισμό για την μεταβλητή Y ώστε να εξομαλύνουμε η και να εξαλείψουμε (κάποιες φορές) τις αποκλίσεις του μοντέλου. χρησιμοποιείτε σε περίπτωσης που διαπιστώνεται έλλειψη κανονικότητας των σφαλμάτων . $Y_t' = \frac{Y^t - 1}{t}$, $t \neq 0$

Διάγραμμα 5.1 *Τα δεδομένα σε λογαριθμική κλίμακα*



Από ότι βλέπουμε από το παραπάνω διάγραμμα το μοντέλο αυτό είναι καλό διότι η παράβλεψη μας είναι αρκετά ακριβής αν και έχει μια αυξητική τάση γενικά ενώ στα πραγματικά δεδομένα δεν ισχύει άρα μπορεί και να μην είναι καλό μοντέλο για μακροχρόνια πρόβλεψη. Από ότι φαίνεται και από τον παρακάτω πινάκα έχει υψηλή ακρίβεια .

Πινάκας 5.1 *Αποτελέσματα εκτίμησης 2006-2016*

Point	Forecast	Test	Lo 80	Hi 80	Lo 95	Hi 95
2006	12.01764	12.00633	11.58762	12.44766	11.35998	12.67530
2007	12.14971	12.10086	11.60775	12.69166	11.32086	12.97855
2008	12.28177	12.17299	11.61947	12.94408	11.26886	13.29468
2009	12.41384	12.25917	11.62333	13.20434	11.20487	13.62281
2010	12.54590	12.29070	11.61982	13.47198	11.12958	13.96222
2011	12.67797	12.37478	11.60931	13.74662	11.04360	14.31233
2012	12.81003	12.42728	11.59215	14.02791	10.94744	14.67262
2013	12.94209	12.47365	11.56860	14.31558	10.84152	15.04266
2014	13.07416	12.51250	11.53894	14.60937	10.72625	15.42207
2015	13.20622	12.53254	11.50339	14.90906	10.60196	15.81048
2016	13.33829	12.55846	11.46214	15.21444	10.46896	16.20761

Στον παραπάνω πινάκα βλέπουμε τα αριθμητικά αποτελέσματα που έχουμε από την πρόβλεψη του μοντέλου καθώς επίσης, όπως και προηγουμένως δίνονται και δυο διαστήματα εμπιστοσύνης (80% και 95 %) που απεικονίζονται και στο διάγραμμα 5.1 με τις το γαλάζιο σκούρο και γαλάζιο ανοιχτό αντίστοιχα 0.3355 – 0.4198

Πίνακας 5.2

	ME	RMSE	MAE	MPE	MAPE	MASE
Training set	0.03795883	0.3355473	0.1969913	0.9201631	5.642610	0.8202836
Test set	-0.34075954	0.4198144	0.3407595	-2.7353415	2.735342	1.4189433

	ACF1	Theil's U
Training set	0.07399341	NA
Test set	0.72255226	7.071223

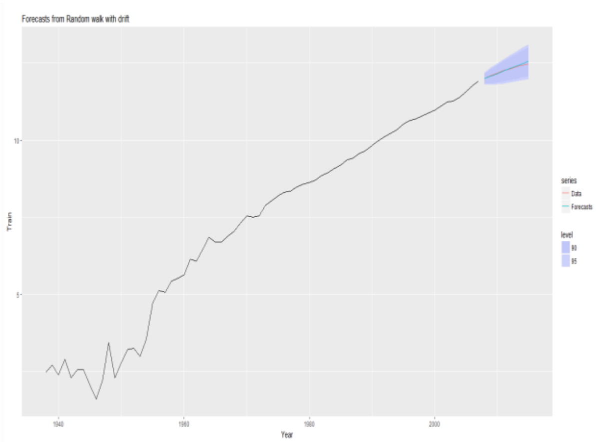
Η απόκλιση της πρόβλεψης από την πραγματική τιμή στην περίοδο πρόβλεψης είναι ικανοποιητική αλλά μεγαλύτερη από το μοντέλο ARIMA που κάναμε προηγούμενος .

6. Υπόδειγμα Τυχαίας Διαδρομής

Σε ένα υπόδειγμα τυχαίας διαδρομής κάθε τιμή της χρονολογικής σειράς, έστω y_t προκύπτει από την αμέσως προηγούμενη της y_{t-1} με την προσθήκη ενός τυχαίου σφάλματος , δηλαδή $y_t = y_{t-1} + \varepsilon_t$ (1), Με ε_t τυχαία σφάλματα . στο υπόδειγμα αυτό παρατηρούμε ότι οι διαδοχικές μεταβολές των τιμών y_t είναι ανεξάρτητες τυχαίες μεταβλητές. Δηλαδή $\Delta_{y_t} = y_t - y_{t-1} = \varepsilon_t$ (2) .Ισχύει επίσης ότι $y_t = y_0 + \sum_{i=1}^{t} \varepsilon_i$ (3) . Η εξίσωση αυτή αποτελεί και τη γενική λύση της εξίσωσης διαφορών πρώτης τάξης που αντιπροσωπεύει η τυχαία διαδρομή (2).

Η σειρά y_tτης τυχαίας διαδρομής (1) έχει σταθερό μέσο καθώς $E(y_t) = y_0$.Η διακύμανση της και οι συνδιακυμάνσεις των τιμών της δεν παραμένουν σταθερές διαχρονικά και άρα η y_t είναι μη στάσιμη .

Διάγραμμα 6.1



Πίνακας 6.1 *Αποτελέσματα εκτίμησης 2008-2015*

Point	Forecast	Test	Lo 90	Hi 90	Lo 95	Hi 95
2008	11.99195	12.00633	11.82539	12.15622	11.79321	12.18744
2009	12.07365	12.10086	11.83738	12.30538	11.79158	12.34928
2010	12.15480	12.17299	11.86474	12.43809	11.80836	12.49163
2011	12.23541	12.25917	11.89982	12.56205	11.83444	12.62366
2012	12.31550	12.29070	11.93963	12.68023	11.86627	12.74891
2013	12.39507	12.37478	11.98267	12.79417	11.90204	12.86922
2014	12.47413	12.42728	12.02805	12.90479	11.94069	12.98567
2015	12.55269	12.47365	12.07517	13.01269	11.98152	13.09897

Από ότι παρατηρούμε από τον παρακάτω πινάκα αποτελεσμάτων της πρόβλεψης που έγινε με την μέθοδο του τυχαίου περιπάτου , Η πρόβλεψη που έγινε είναι αρκετά ακριβής σε σχέση με όλες της προηγούμενες που έχουν γίνει έως τώρα κάτι που φαίνεται και από το διάγραμμα 6.1 .

Πίνακας 6.2

```
Forecast method: Random walk with drift

Drift: 0.9831  (se 0.1431)
Residual sd: 1.1973

Error measures:
                      ME       RMSE       MAE         MPE
Training set -0.04176586 0.33285629 0.1851199 -2.92576427
Test set     -0.01092849 0.03760038 0.0318130 -0.08605235
                 MAPE      MASE        ACF1
Training set 5.7706248 0.7708504 -0.01023242
Test set     0.2580168 0.1324712  0.53280579
```

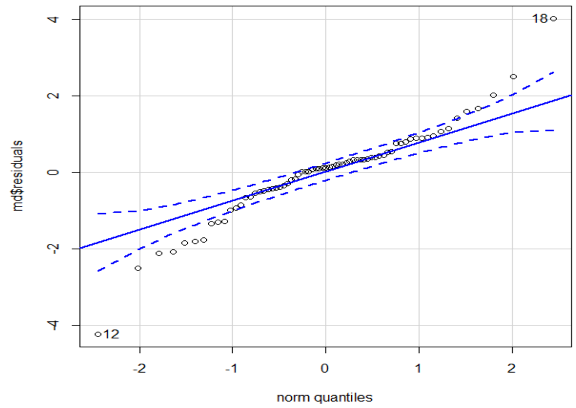
Το υπόδειγμα του τυχαίου περιπάτου που εκτελέσαμε μας έβγαλε και μια σταθερά (drift=0.9831) . Αν ένα υπόδειγμα της τυχαίας διαδρομής περιλαμβάνει και τον σταθερό ορό αυτόν τότε έχουμε υπόδειγμα τυχαίας διαδρομής με σταθερά . Όπου β (drift) είναι η σταθερά της παρακάτω εξίσωσης .

$\Delta_{y_t} = \beta + \varepsilon_t$ *(4)*

Είναι σημαντικό να αναφέρουμε ότι ,σε σύγκριση με το υπόδειγμα τυχαίας διαδρομής , οι πρώτες διαφορές στο υπόδειγμα με σταθερά είναι εν μέρει στοχαστικές και εν μέρει σταθερές .

$y_t = y_0 + \beta t + \sum_{i=1}^{t} \varepsilon_i$ *(5)*

Διάγραμμα 4.2 *qqplot καταλοίπων*



Παρατηρούμε ότι και σε αυτό το μοντέλο τα σφάλματα έχουν πολύ έντονα παχιές ουρές κάτι το όποιο αναμένετε γιατί στις περισσότερες χρονολογικές σειρές παρατηρείτε αυτό το φαινόμενο, και θυμίζουν περισσότερο ότι ακλουθούνε την t-student κατανομή και όχι κανονική.

7. Αποτελέσματα Μοντέλων

Στο παρακάτω πινάκα βλέπουμε τα αποτελέσματα του RMSE για τα 3 μοντέλα και διαπιστώνουμε ότι το μοντέλο με την καλύτερη ικανότητα πρόβλεψης για το μέλλον είναι το υπόδειγμα τυχαίας διαδρομής με σφάλμα για το test set να είναι 0.03760 αρκετά μικρότερο από όλα τα αλλά υποδείγματα .

Πίνακας 7.1

Υπόδειγμα	***ARIMA***	***Εκθετική εξομάλυνση***	***Τυχαίας διαδρομής***
Train Set	0.2835	*0.3355*	*0.3328*
Test set	0.0738	*0.4198*	*0.03760*

8. Εξαγωγή δεδομένων

Στο ερώτημα που έχει γίνει στο ένα τα δεδομένα ήταν πολλαπλασιασμένα επί (10,100,1000,100000,…) μια λύση θα ήταν εφικτή για την ανάλυση των δεδομένων είναι να χρησιμοποιηθεί κάποιο εργαλείο για Big data όπως για παράδειγμα το Spark το όποιο είναι κατάλληλο για την ανάλυση μεγάλου όγκου δεδομένων. Τα δεδομένα για αυτήν την εργασία τα διαβάσαμε με το command line ενός VM ubundu server όπως έχουμε αναφέρει και στο πρώτο κεφάλαιο με χαρακτηριστικά 4GB Ram x1 Cpu και 20 GB χωρητικότητα . Έγιναν πειράματα να αναπαράγουμε 2 φορές τα δεδομένα με ένα copy-paste από το command line και να ξανά κάνουμε την ιδία διαδικασία ώστε να δούμε την διάφορα του χρόνου στα 55 σχεδόν εκατομμύρια με σύγκριση τα 111 *εκατομμύρια δεδομένα. Το ίδιο πείραμα προσπαθήσαμε να το κάνουμε και για 4 φορές τα δεδομένα μας αλλά το μηχάνημα δεν τα κατάφερε να το κάνει .Τα κανονικά δεδομένα έκανε να τα ανοίξει 8.40 δευτερόλεπτα ενώ τα διπλασιαζόμενα δεδομένα έκανε 1.30 λεπτά ,άρα καταλαβαίνουμε ότι υπάρχει μεγάλη διάφορα . Επόμενος σε παράλληλο σύστημα όπως το Spark θα ήταν μια πολύ καλή λύση.*

Βιβλιογραφία

[1]. *Δημελη Σ. ; (2003) ; Σύγχρονες μέθοδοι ανάλυσης χρονολογικών σειρών.*
[2]. *Μαρκοπουλος Α. , Ντεντης Ι. Παρασκευοπουλος Ν ; (2015) ; Η χρήση της μεθοδολογίας box-jenkins στην ανάλυση χρονοσειρων.*
[3]. *Εμιρης Δ. ; (2012) ; Προβλέψεις .*
[4]. *Κουντουρη Φ. ; (2008) ; Χρονολογικες σειρες.*
[5]. *Γιώργος Θεοδώρου , Δημήτρης Κουγιουμτζής ; (2009) ; Μοντέλων Χρονοσειρών και Πρόβλεψη.*
[6]. *Σαριαννίδης Ν. ; Οικονομετρία ; Ανάλυση Χρονολογικών Σειρών ;*
[7]. *Τεχνικές προβλέψεων & έλεγχου; Μάθημα θεωρίας στάσιμες διαδικασίες - υποδείγματα .*