



Σκοπός

Σκοπός της παρούσας εργασίας είναι η παρουσίαση των σημαντικότερων τεχνικών πρόβλεψης χρονοσειρών που περιλαμβάνουν τις παραδοσιακές μεθόδους πρόβλεψης (μοντέλα ARIMA), τα τεχνητά νευρωνικά μοντέλα (ANN) και τα υβριδικά μοντέλα και η πρόταση ενός τρόπου επιλογής του καλύτερου μοντέλου πρόβλεψης με βάση ορισμένα χαρακτηριστικά των πραγματικών συνόλων δεδομένων.

Χρονοσειρές

Ως χρονοσειρά (time series) ορίζεται μια σειρά από παρατηρήσεις $x_1, x_2, \dots, x_t$ ενός φαινομένου οι οποίες παρατηρούνται σε συγκεκριμένες χρονικές στιγμές που ισαπέχουν μεταξύ τους (π.χ. ανά έτος, τρίμηνο, ημέρα) και εκφράζουν την εξέλιξη ενός φαινομένου στο χρόνο. Οι παρατηρήσεις $x_1, x_2, \dots, x_t$ είναι συγκεκριμένες τιμές των τυχαίων μεταβλητών $X_1, X_2, \dots, X_t$ και είναι μέρος μια άπειρης ακολουθίας τυχαίων αριθμών. Η ακολουθία αυτή ονομάζεται στοχαστική διαδικασία και συμβολίζεται με $\{X_t\}_{t \geq 0}$.

Βασικά Στατιστικά Μεγέθη

Μέση Τιμή ή Ροπή $1^{ης}$ τάξης

$$E(X_t) = \int x \cdot f(x) dx$$

Αυτοσυνδιακύμανση (autocovariance)

$$\gamma_h = cov(X_t, X_{t+h}) = E[X_t X_{t+h}] - \mu_t \mu_{t+h}$$

όπου X_t και X_{t+h} δύο τιμές της ίδιας χρονολογικής σειράς τις t και $t+h$ στιγμές και $E(X_t) = \mu_t$

Διασπορά

Για $h=0$, η αυτοσυνδιακύμανση ταυτίζεται με τη διακύμανση δηλαδή

$$\gamma_0 = Var(X_t) = \sigma^2 = E[(X_t - \mu_t)^2]$$

Αυτοσυσχέτιση (autocorrelation)

$$\rho_h = \frac{\gamma_h}{\gamma_0}, \quad h = 1, 2, \dots$$

Χαρακτηριστικά Χρονοσειρών

Τάση: Η συστηματική μεταβολή (γραμμική ή μη) των τιμών ενός χαρακτηριστικού στη μονάδα του χρόνου.

Εποχικότητα: Οι τιμές της χρονολογικής σειράς τείνουν να εμφανίζονται αυξημένες ή μειωμένες σε συγκεκριμένες χρονικές περιόδους

Στασιμότητα: Οι στατιστικές ιδιότητες της χρονοσειράς παραμένουν αναλλοίωτες καθ' όλη τη διάρκεια της υπό μελέτη χρονικής περιόδου.

Εποχικό ολοκληρωμένο αυτοπαλινδρομούμενο

εποχικό μοντέλο SARIMA(p, d, q)x(P, D, Q)_s

Στην παρούσα εργασία χρησιμοποιείται το εποχικό πρότυπο, αφού τα δεδομένα παρουσιάζουν τάση και εποχικότητα.

$$\varphi(B)[\Phi(B^s)(1-B^s)^D](1-B)^d y_t = \theta(B)\Theta(B^s)Z_t$$

όπου $Z_t \sim WN(0, \sigma_z^2)$

Μεθοδολογία Box-Jenkins

1. Ταυτοποίηση υποδείγματος

Προσδιορισμός της τάξης μη στασιμότητας και των AR και MA όρων του μοντέλου

2. Εκτίμηση του υποδείγματος

Εκτίμηση των παραμέτρων του μοντέλου.

3. Έλεγχος του υποδείγματος

Ελέγχεται αν τα υπόλοιπα (residuals) είναι λευκός θόρυβος και γίνεται η επιλογή του μοντέλου με βάση τα κριτήρια BIC και AIC.

4. Πρόβλεψη

Πραγματοποιείται η πρόβλεψη των μελλοντικών τιμών της χρονοσειράς και υπολογίζεται το σφάλμα πρόβλεψης (prediction error).

Μοντελοποίηση χρονοσειρών με τεχνητά νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα αποτελούν ένα δίκτυο νευρώνων σε αντιστοιχία με τη λειτουργία του ανθρώπινου εγκεφάλου. Ένας τεχνητός νευρώνας αποτελείται από τα σήματα εισόδου (input layer) $x_1, x_2, \dots, x_m$. Κάθε είσοδος πολλαπλασιάζεται με το ένα συντελεστή w_{jk} ο οποίος ονομάζεται συναπτικό βάρος. Το k συμβολίζει το νευρώνα του δικτύου. Όλες οι εισοδοί πολλαπλασιάζονται με τα συναπτικά τους βάρη και στη συνέχεια αθροίζονται σε έναν γραμμικό αθροιστή (linear combiner). Στο άθροισμα προστίθεται και η πόλωση b_j που έχει τιμή εισόδου ίση με τη μονάδα. Το συνολικό αυτό άθροισμα ονομάζεται τοπικό πεδίο (net input) και συμβολίζεται με v_k . Το τοπικό πεδίο εισέρχεται σε μία συνάρτηση η οποία ονομάζεται συνάρτηση ενεργοποίησης (activation function) και έτσι προκύπτει η έξοδος y_k του νευρώνα k . Η συνάρτηση ενεργοποίησης είναι συνήθως η σιγμοειδής συνάρτηση. Τα βασικά παραδείγματα σιγμοειδών συναρτήσεων είναι η λογιστική συνάρτηση (logistic function) και η υπερβολική εφαπτομένη (hyperbolic tangent) η οποία χρησιμοποιείται και στην παρούσα εργασία.

Αρχιτεκτονική Δικτύου

Στη μοντελοποίηση και πρόβλεψη χρονοσειρών το πιο διαδεδομένο δίκτυο είναι το δίκτυο ενός κρυφού επιπέδου και εμπρόσθιας τροφοδότησης (Single Hidden Layer Feedforward Network). Η σχέση μεταξύ εισόδου και εξόδου δίνεται από τον τύπο:

$$y_t = w_o + \sum_{j=1}^q w_j g(w_{oj} + \sum_{i=1}^p w_{ij} y_{t-1}) + v_t$$

όπου w_j και w_{ij} : συναπτικά βάρη, p : κόμβοι εισόδου, q :

πλήθος κρυφών επιπέδων, g : συνάρτηση ενεργοποίησης

Για την πρόβλεψη είναι σημαντικό να καθοριστεί ο αριθμός των κόμβων εισόδου (input layer). Εκτός από το επίπεδο εισόδου, πρέπει να ορισθεί ο αλγόριθμος μάθησης, ο οποίος στην παρούσα εργασία είναι ο αλγόριθμος ανάστροφης διάδοσης του σφάλματος (error backpropagation algorithm). Τέλος, πρέπει να καταλήξουμε στο βέλτιστο αριθμό κόμβων στο κρυφό επίπεδο. Η συνάρτηση ενεργοποίησης του κάθε νευρώνα είναι η υπερβολική εφαπτομένη, ενώ το δίκτυο δίνει μια τιμή απόκρισης, δηλαδή κάνει πρόβλεψη 1-βήματος μπροστά.

Υβριδικά Μοντέλα

Τα υβριδικά μοντέλα βασίζονται στην υπόθεση ότι η χρονοσειρά αποτελείται από τη γραμμική και τη μη γραμμική συνιστώσα δηλαδή:

$$y_t = L_t + N_t$$

Αρχικά με το μοντέλο SARIMA μοντελοποιούμε τη γραμμική συνιστώσα. Αν e_t είναι το υπόλοιπο τη στιγμή t από το γραμμικό μοντέλο, τότε

$$e_t = y_t - \hat{L}_t$$

όπου $\hat{L}_t$: η προβλεπόμενη τιμή για τη χρονική στιγμή t από το γραμμικό μοντέλο SARIMA

Για να εντοπίσουμε τις μη γραμμικότητες μεταξύ των υπολοίπων χρησιμοποιούμε τα τεχνητά νευρωνικά δίκτυα. Τα υπόλοιπα του γραμμικού μοντέλου εισάγονται στο δίκτυο

$$e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-A}) + v_t$$

όπου f : μη γραμμική συνάρτηση των κόμβων του δικτύου, v_t :

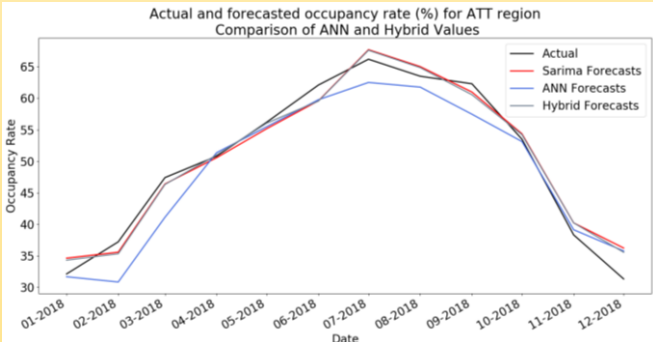
θόρυβος, A : αριθμός των κόμβων εισόδου

Οι προβλέψεις συνεπώς του υβριδικού μοντέλου προκύπτουν από τις προβλέψεις του γραμμικού μοντέλου SARIMA και του μη γραμμικού μοντέλου ANN δηλαδή:

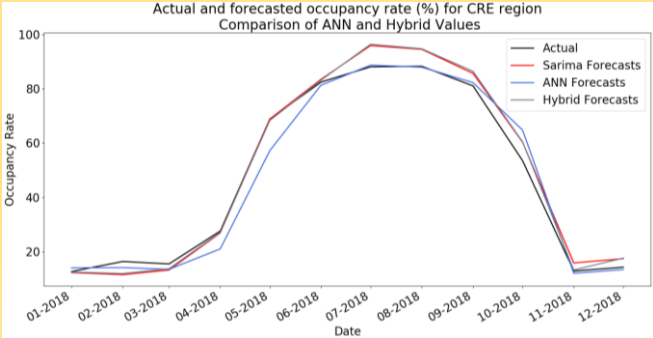
$$\hat{y}_t = \hat{L}_t + \hat{N}_t$$

Εφαρμογή

Μελετήσαμε πραγματικά σύνολα δεδομένων που συλλέχθηκαν από την ΕΛΣΤΑΤ και αφορούν στην πληρότητα των ξενοδοχειακών μονάδων των 13 περιφερειών της Ελλάδας από το 2006 έως το 2018. Πραγματοποιήσαμε προβλέψεις των τελευταίων 12 μηνών με τα βέλτιστα μοντέλα SARIMA, τα νευρωνικά δίκτυα και τα υβριδικά μοντέλα για κάθε περιφέρεια. Ενδεικτικά στο Σχήμα 1 και στο Σχήμα 2 φαίνονται οι προβλέψεις για την περιφέρεια της Αττικής και της Κρήτης αντίστοιχα.



Σχήμα 1: Σύγκριση πραγματικών και εκτιμώμενων τιμών από το βέλτιστο μοντέλο SARIMA, το βέλτιστο ANN μοντέλο και το βέλτιστο υβριδικό μοντέλο για την περιφέρεια της Αττικής (ATT)



Σχήμα 2: Σύγκριση πραγματικών και εκτιμώμενων τιμών από το βέλτιστο μοντέλο SARIMA, το βέλτιστο ANN μοντέλο και το βέλτιστο υβριδικό μοντέλο για την περιφέρεια της Κρήτης (CRE)

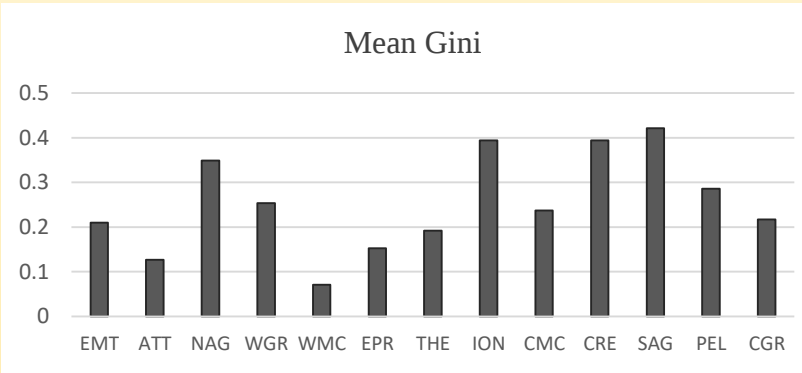
Από τα διαγράμματα φαίνεται ότι η περιφέρεια της Κρήτης πετυχαίνει καλύτερες προβλέψεις με τα ANN, ενώ η περιφέρεια της Αττικής πετυχαίνει καλύτερες προβλέψεις με τα Υβριδικά Μοντέλα.

Παρατηρήσαμε ότι η ομαδοποίηση των περιφερειών με βάση τα καλύτερα μοντέλα πρόβλεψης σχετίζεται με την εποχικότητα. Ένας δείκτης μέτρησης της εποχικότητας είναι ο συντελεστής Gini ο οποίος ορίζεται ως

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{2n^2 \bar{x}}$$

όπου n : σύνολο δεδομένων, x_i : παρατήρηση i , x_j : παρατήρηση j , $\bar{x}$: μέση τιμή

Παίρνει τιμές από το 0 έως το 1. Όσο πιο κοντά στο 0 τόσο μικρότερη είναι η εποχικότητα. Στο Σχήμα 3 φαίνεται ο μέσος συντελεστής Gini ανά περιφέρεια.



Σχήμα 3: Μέσος συντελεστής Gini ανά περιφέρεια

Ορισμένες περιφέρειες της Ελλάδας όπως τα νησιά και η Πελοπόννησος εμφανίζουν υψηλούς μέσους συντελεστές Gini, ενώ οι υπόλοιπες περιφέρειες εμφανίζουν σημαντικά χαμηλότερες τιμές όπως η Δυτική Μακεδονία και η Αττική. Παρατηρήσαμε ότι οι περιφέρειες που εμφανίζουν τους υψηλούς συντελεστές Gini πραγματοποιούν καλύτερες προβλέψεις με τα ANN, ενώ οι περιφέρειες που εμφανίζουν χαμηλούς συντελεστές Gini πετυχαίνουν καλύτερες προβλέψεις με τα υβριδικά μοντέλα. Συνεπώς, ο συντελεστής Gini που είναι το μέτρο της εποχικότητας μπορεί να αποτελέσει το χαρακτηριστικό με βάση το οποίο θα μπορούσε να επιτευχθεί η ομαδοποίηση.

Χρησιμοποιώντας τη μέθοδο K-means πραγματοποιούμε την ομαδοποίηση των περιφερειών στις δύο κατηγορίες και στη συνέχεια χρησιμοποιώντας την καμπύλη ROC και τον δείκτη Youden βρίσκουμε το βέλτιστο σημείο αποκοπής (cutoff point) που αντιστοιχεί σε τιμή Gini ίση με 0.28.

Συμπεράσματα

Στην παρούσα διπλωματική εργασία παρουσιάστηκαν οι βασικότερες τεχνικές πρόβλεψης που περιλαμβάνουν τις παραδοσιακές μεθόδους πρόβλεψης (μοντέλα ARIMA/SARIMA), τα Τεχνητά Νευρωνικά Δίκτυα (ANN) και τα Υβριδικά Μοντέλα. Είδαμε ότι δεν υπάρχει ένα ενιαίο βέλτιστο μοντέλο πρόβλεψης για όλες τις περιφέρειες και προτείναμε έναν τρόπο επιλογής του καλύτερου μοντέλου πρόβλεψης των χρονοσειρών με βάση το συντελεστή Gini, ο οποίος είναι δείκτης εποχικότητας των χρονοσειρών. Με χρήση της καμπύλης ROC και του δείκτη Youden βρήκαμε το βέλτιστο σημείο αποκοπής και προτείναμε ότι για χρονοσειρές με έντονη εποχικότητα (δείκτης $Gini \geq 0.28$) τα ANN αποτελούν καλύτερο μοντέλο πρόβλεψης και για χρονοσειρές με μικρή εποχικότητα (δείκτης $Gini < 0.28$) τα υβριδικά μοντέλα αποτελούν καλύτερο μοντέλο πρόβλεψης.