ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάκτηση δεδομένων που λείπουν σχετικών με τις συνθήκες άνεσης ενός κτηρίου και πρόβλεψη της παρουσίας ενοίκων σε αυτό.

Διπλωματική εργασία

Βασίλειος Σούλτης

Α.Μ: 2006030083, Email: vsoultis@isc.tuc.gr

Εξεταστική Επιτροπή:

Κ. Καλαϊτζάκης, Καθηγητής, Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών, Πολυτεχνείο Κρήτης (Επιβλέπων)

Δ. Ρόβας, Επίκουρος Καθηγητής, Τμήμα Μηχανικών Παραγωγής και Διοίκησης, Πολυτεχνείο Κρήτης

Ε. Κουτρούλης, Επίκουρος Καθηγητής, Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών, Πολυτεχνείο Κρήτης

Ανάκτηση δεδομένων που λείπουν σχετικών με τις συνθήκες άνεσης ενός κτηρίου και πρόβλεψη της παρουσίας ενοίκων σε αυτό.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Περίληψη

Στον συνεχώς αναπτυσσόμενο τομέα διαχείρισης των ενεργειακών πόρων με σκοπό την μείωση κατανάλωσης ενέργειας του πλανήτη, η ενεργειακή διαχείριση κτηρίων συγκαταλέγεται ανάμεσα στους πιο σημαντικούς τομείς γύρω από τους οποίους κινείται η διεθνής έρευνα και αναπτύσσονται συνεχώς νέες εφαρμογές.

Η παρούσα διπλωματική εργασία μελετά μεθόδους ανάκτησης δεδομένων των συνθηκών ενός κτηρίου καθώς επίσης και μεθόδους πρόβλεψης της παρουσίας ενοίκων στο κτήριο. Στόχος των μεθόδων αυτών είναι να διασφαλίσουν την ακεραιότητα των δεδομένων, με σκοπό την ορθότερη αξιολόγηση τους και μετέπειτα αξιοποίηση τους στα συστήματα διαχείρισης του κτηρίου, καθώς και να προσδώσουν στην λογική λειτουργίας των συστημάτων την ικανότητα πρόβλεψης παραμέτρων όπως είναι η παρουσία ενοίκων, εκτιμώντας τις ώρες που θα πρέπει οι χώροι του κτηρίου να κλιματίζονται.

Πιο συγκεκριμένα αφού αναπτύχθηκε αναλυτικά το θεωρητικό υπόβαθρο των μεθόδων που μελετήθηκαν, πραγματοποιήθηκαν πειράματα χρήσης των μεθόδων χρησιμοποιώντας πραγματικά δεδομένα θερμοκρασίας χώρου και παρουσίας ατόμων και εξήλθαν συμπεράσματα σχετικά με την αποτελεσματικότητα των μεθόδων ανάλογα με τα χαρακτηριστικά των δειγμάτων που χρησιμοποιήθηκαν.

Κεφάλαιο 1

Εισαγωγή

Στη σύγχρονη εποχή παρατηρείται ραγδαία εξέλιξη των τεχνολογικών επιστημών. Μεγάλη προσπάθεια καταβάλλεται στην βελτιστοποίηση των υπηρεσιών και προϊόντων προς τους πολίτες, διευκολύνοντας την καθημερινότητά τους. Η εξέλιξη αυτή συμβάλλει σημαντικά και στις απαιτήσεις του σύγχρονου κόσμου για μείωση της ενέργειας που καταναλώνεται στην καθημερινότητα των ανθρώπων, με απώτερο σκοπό την μείωση της ρύπανσης του περιβάλλοντος και βέβαια στις μέρες μας στην βελτίωση της οικονομίας των καταναλωτών. Με άλλα λόγια παρατηρείται μία συνεχής προσπάθεια αύξησης της αποδοτικότητας, σε όλους τους τομείς της σύγχρονης ζωής, μέσω αύξησης των επιδόσεων και παράλληλα μείωση της καταναλισκόμενης ενέργειας. Ισχυρά παραδείγματα μέσα από την καθημερινότητα των πολιτών προς αυτή την κατεύθυνση αποτελούν ο τομέας των οχημάτων και ο τομέας των οικιακών συσκευών, όπου παρατηρείται έντονη προσπάθεια μείωσης της κατανάλωσης καυσίμου και ηλεκτρικού ρεύματος αντίστοιχα με παράλληλη βελτίωση των επιδόσεων τους.

Η εξέλιξη αυτή δεν θα μπορούσε να αφήσει ανεπηρέαστο τον κτηριακό τομέα καθώς αποτελεί μείζον κομμάτι της καθημερινής ζωής του ανθρώπου με αποτέλεσμα τα επίπεδα της καταναλισκόμενης ενέργειας των κτηρίων να είναι αρκετά υψηλά. Έτσι κρίνεται επιτακτική η ανάγκη ενασχόλησης με τον συγκεκριμένο τομέα με στόχο την ανάπτυξη τεχνολογιών για την μείωση της ενέργειας που καταναλώνουν τα κτήρια αλλά και την περαιτέρω βελτίωση των κτηριακών συνθηκών.

Η εξοικονόμηση ενέργειας σε ένα κτήριο εξασφαλίζεται εν μέρει με τον κατάλληλο σχεδιασμό του κτηρίου και τη χρήση ενεργειακά αποδοτικών δομικών στοιχείων και συστημάτων και εν μέρει μέσω της υψηλής αποδοτικότητας των εγκατεστημένων συστημάτων κλιματισμού και θέρμανσης, η οποία προϋποθέτει την άριστη ποιότητα του σχετικού εξοπλισμού και της εγκατάστασής του. Άλλος ένας καθοριστικός παράγοντας εξοικονόμησης ενέργειας είναι η σωστή διαχείριση του κτηρίου, όσο αφορά τον εξοπλισμό του με απώτερο σκοπό την μείωση της καταναλισκόμενης ενέργειας. Η σωστή διαχείριση του κτηρίου στοχεύει στην εξασφάλιση συνθηκών και υπηρεσιών τέτοιων που να κάνουν την παραμονή των ενοίκων στα κτήρια ευχάριστη με την ελάχιστη δυνατή ενεργειακή κατανάλωση.

Η ενεργειακή διαχείριση των κτηρίων γίνεται ως επί το πλείστον από τους ίδιους τους ενοίκους των κτηρίων. Αυτό έχει σαν αποτέλεσμα η ενεργειακή απόδοση ενός κτηρίου καθώς και η λειτουργία των ενεργειακών συστημάτων του να καθορίζεται πλήρως από την ενεργειακή συμπεριφορά του χρήστη. Έτσι από την μη ορθολογική χρήση του κτηρίου και των συστημάτων του, μπορεί να μειωθεί σημαντικά η ενεργειακή απόδοση ενός κτηρίου.

Για την ελαχιστοποίηση της επίδρασης του ενοίκου στην ενεργειακή απόδοση του κτηρίου κρί-

νεται επιτακτική η ανάγκη εγκατάστασης και χρήσης συστημάτων ελέγχου και αυτοματισμού. Μέσω αυτών των συστημάτων αποφεύγονται καταστάσεις μη ορθολογικής χρήσης των συστημάτων του κτηρίου. Για το λόγο αυτό αναπτύσσονται συνεχώς και διάφοροι τρόποι για τον έλεγχο και την διαχείριση των συσκευών και συστημάτων των κτηρίων με σκοπό να βελτιστοποιήσουν την ενεργειακή τους κατανάλωση διατηρώντας ωστόσο την άνεση των ανθρώπων εντός των κτηρίων.

Σημαντικός παράγοντας για την ενίσχυση των μεθόδων ελέγχου και διαχείρισης των κτηρίων είναι η δυνατότητα σωστής εκμετάλλευσης των δεδομένων που επιστρέφουν τα συστήματα ελέγχου. Η συγκεκριμένη ενέργεια αποσκοπεί στην σωστότερη εκτίμηση της κατάστασης του κτηρίου με αποτέλεσμα την καλύτερη διαχείριση των ενεργειακών του συστημάτων ώστε να βελτιστοποιηθεί από την μία η ενεργειακή απόδοση και από την άλλη να παραμείνει σε υψηλά επίπεδα η άνεση των ενοίκων του κτηρίου.

1.1 Αντικείμενο της Διπλωματικής Εργασίας

Πολλές φορές τα δεδομένα που επιστρέφουν τα συστήματα ελέγχου ενός κτηρίου είναι είτε κατεστραμμένα είτε ελλειπή. Αυτό οφείλεται σε διάφορους λόγους όπως είναι η δυσλειτουργία κάποιων από τους αισθητήρες του συστήματος ελέγχου, στην διακοπή παροχής ηλεκτρικού ρεύματος για κάποιο χρονικό διάστημα και άλλα. Έτσι μην έχοντας σαφή απεικόνιση των δεδομένων του κτηρίου όπως θερμοκρασία, υγρασία κτλ, δεν εκτιμάται σωστά η γενικότερη κατάσταση του κτηρίου με αποτέλεσμα να μην καθίσταται αποτελεσματική η ενεργειακή διαχείρισή του. Ουσιαστικά η έλλειψη δεδομένων ενός κτηρίου οδηγεί στην λάθος εκτίμηση των χαρακτηριστικών του που προσδιορίζονται από τα συγκεκριμένα δεδομένα, με αποτέλεσμα την εσφαλμένη αξιολόγηση της γενικότερης συμπεριφοράς του κτηρίου που έχει να κάνει με την διαχείριση της ενέργειας που καταναλώνει αλλά και με την αποτελεσματικότητα του κτηρίου σχετικά με τις συνθήκες άνεσης που προσφέρει. Αντικείμενο λοιπόν της παρούσας διπλωματικής εργασίας είναι αρχικά η παράθεση μεθόδων οι οποίες καθιστούν δυνατή την διόρθωση δεδομένων. Πιο συγκεκριμένα, γίνεται μια προσπάθεια ανάλυσης του τρόπου λειτουργίας μεθόδων παλινδρόμησης (regression) ή προσαρμογής καμπύλης (curve fitting) και εφαρμογή συγκεκριμένων πειραμάτων με σκοπό την αποτίμηση της αποδοτικότητας των μεθόδων ανάλογα με τις περιπτώσεις των δεδομένων που επεξεργάζονται σε κάθε πείραμα.

Σημαντικά στοιχεία που ενισχύουν την ενεργειακή διαχείριση των κτηρίων, εκτός από την αδιάκοπη και ορθή πληροφόρηση των συστημάτων ελέγχου του κτηρίου μέσω των δεδομένων του, είναι και η γνώση παραμέτρων όπως των χρονικών περιόδων μέσα στην ημέρα όπου παρατηρείται παρουσία ή απουσία των ενοίκων ενός κτηρίου. Πιο συγκεκριμένα, αυτό που ίσως είναι πιο σημαντικό σχετικά με αυτού του είδους τις παραμέτρους, είναι η δυνατότητα πρόβλεψης των τιμών τους ώστε το κτήριο να προετοιμάζεται και να προσαρμόζεται προλαμβάνοντας τις ανάγκες λειτουργίας των συστημάτων του που προβλέπεται να προκύψουν, με στόχο να καλυφθούν οι απαιτήσεις των ενοίκων. Για παράδειγμα κάνοντας πρόβλεψη της πληρότητας ενός κτηρίου, επιτυγχάνεται ορθολογικότερη χρήση των συστημάτων του που πρακτικά απενεργοποιούνται με την απουσία των ενοίκων ή ενεργοποιούνται λίγο πριν την προβλεπόμενη παρουσία ενοίκων με σκοπό να μην παρατηρείται απώλεια άνεσης των ενοίκων αλλά και να επιτυγχάνονται σημαντικά οφέλη, οικονομικά και περιβαλλοντολογικά, καθώς μειώνεται δραστικά η ενέργεια που καταναλώνεται. Έτσι λοιπόν με αυτό τον τρόπο δίνεται η δυνατότητα ελέγχου του κτηρίου με τέτοιο τρόπο ώστε να συγχρονίζεται η λειτουργία των συστημάτων

του με το πρόγραμμα κατοίκησης του από τους ενοίκους, δηλαδή με το πότε βρίσκονται εντός του κτηρίου και το πότε φεύγουν από αυτό. Για το λόγο αυτό, ένα μέρος της παρούσας διπλωματικής εργασίας ασχολείται με την ανάπτυξη μεθόδων που να μπορούν να προβλέπουν τις μελλοντικές τιμές τέτοιου είδους παραμέτρων όσο πιο αποτελεσματικά γίνεται χρησιμοποιώντας δεδομένα που έχουν καταχωρηθεί σε παλαιότερα χρονικά διαστήματα.

1.2 Οργάνωση της Εργασίας

Η παρούσα διπλωματική εργασία αναπτύχθηκε σε συνολικά έξι κεφάλαια. Στο τρέχον κεφάλαιο, το πρώτο, γίνεται η εισαγωγή στην εργασία. Στο δεύτερο κεφάλαιο παρουσιάζονται οι εργασίες που θέτουν το πλαίσιο στο οποίο κινείται η παρούσα διπλωματική εργασία. Το τρίτο κεφάλαιο ασχολείται αναλυτικά με την θεωρία των μεθόδων που μελετήθηκαν στην παρούσα διπλωματική εργασία τόσο από την μεριά της διόρθωσης δεδομένων όσο και από την μεριά της πρόβλεψης πληρότητας. Στο τέταρτο κεφάλαιο παρουσιάζονται πειράματα των μεθόδων διόρθωσης δεδομένων και αντίστοιχα στο πέμπτο κεφάλαιο παρουσιάζονται πειράματα των μεθόδων πρόβλεψης πληρότητας. Τέλος στο έκτο κεφάλαιο εξάγονται συμπεράσματα για την κάθε μέθοδο ξεχωριστά.

Κεφάλαιο 2

Βιβλιογραφική Ανασκόπηση

Σε αυτό το κεφάλαιο παρουσιάζονται εργασίες που θέτουν το πλαίσιο στο οποίο κινείται η παρούσα διπλωματική εργασία.

2.1 Διόρθωση Δεδομένων

Τις τελευταίες δεκαετίες παρατηρείται μια συνεχής προσπάθεια της επιστήμης να αντιμετωπίσει το πρόβλημα των ελλειπόντων δεδομένων (missing data). Η πολυετής αυτή έρευνα αποδεικνύει το γεγονός ότι είναι πολύ σημαντικό σωστά, αδιάκοπα και ολοκληρωμένα δεδομένα να τίθενται στην διάθεση διαφόρων επιστημονικών πεδίων ούτως ώστε να προκύπτουν πληροφορίες και συμπεράσματα με τη μέγιστη δυνατή ακρίβεια. Το πρόβλημα των ελλειπόντων δεδομένων είναι κοινό για κάθε είδους μελέτη που στηρίζεται σε δεδομένα που συλλέγονται από τον πραγματικό κόσμο όπως για παράδειγμα σε δημοσκοπήσεις, έρευνες αγοράς, ιατρικές μελέτες, έρευνες και εφαρμογές της μηχανικής, σε τομείς επικοινωνιών αλλά και σε χρηματοοικονομικούς τομείς. Ανάλογα το επιστημονικό πεδίο, το πρόβλημα των δεδομένων που λείπουν μπορεί να παρουσιαστεί λόγω κάποιου σφάλματος ενός αισθητήρα, σε πρόβλημα κατά τη μεταφορά δεδομένων ενός ψηφιακού συστήματος ή ακόμη στην παράλειψη μίας ερώτησης του ερωτηματολογίου μίας έρευνας [10, 35, 37, 52]. Η ποιότητα των δεδομένων, δηλαδή σωστά και αδιάκοπα σύνολα δεδομένων, είναι εξίσου σημαντική και στις κτηριακές εφαρμογές που απασχολεί την τρέχουσα διπλωματική εργασία έτσι ώστε να επιτυγχάνεται όσο το δυνατόν βέλτιστη διαχείριση των ενεργειακών συστημάτων ενός κτηρίου με σκοπό να καθίσταται το κτήριο ενεργειακά αποδοτικότερο και παράλληλα να παρέχει την μέγιστη άνεση στους ενοίκους.

Για την καλύτερη κατανόηση και άρα καλύτερη αντιμετώπιση των δεδομένων που λείπουν πραγματοποιείται μια κατηγοριοποίηση αυτών. Έτσι σύμφωνα με τις [5, 32, 35] υπάρχουν τρεις κατηγορίες δεδομένων που λείπουν:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing not at Random (MNAR)

Για να διευκολυνθεί η κατανόηση των παραπάνω κατηγοριών συμβολίζεται ως Z μία μεταβλητή με ελλείποντα δεδομένα και X ένα σύνολο άλλων μεταβλητών όπου τα δεδομένα είναι συνεχώς παρα-

τηρούμενα, δηλαδή δεν έχουν ελλείπουσες τιμές. Έτσι στην κατηγορία MCAR οι τιμές που λείπουν σε μία μεταβλητή (Z) δεν έχουν καμία σχέση με τις τιμές άλλων μεταβλητών (X). Δηλαδή η πιθανότητα του να λείπουν δεδομένα είναι ίδια για όλες τις περιπτώσεις. Αυτό σημαίνει ότι οι αιτίες που οδηγούν σε ελλείποντα δεδομένα είναι άσχετες με οποιαδήποτε δεδομένα, είτε αυτών καθ' αυτών των ελλειπόντων δεδομένων δηλαδή της μεταβλητής Z, είτε των παρατηρούμενων δεδομένων δηλαδή της μεταβλητής X. Για παράδειγμα στην συγκεκριμένη κατηγορία, θεωρώντας έναν αισθητήρα S, η πιθανότητα του να λείπουν δεδομένα από τις μετρήσεις του S δεν εξαρτάται από δεδομένα που μπορεί να υπάρχουν σε μία βάση δεδομένων αλλά δεν εξαρτάται ούτε και από την κατάσταση του ίδιου του αισθητήρα ή κάποιου άλλου αισθητήρα.

Στην επόμενη κατηγορία MAR, η αιτία των ελλειπόντων δεδομένων μπορεί να σχετίζεται με τις παρατηρούμενες μεταβλητές (X) αλλά δεν έχει καμία σχέση με τις μεταβλητές των ίδιων ελλειπόντων δεδομένων δηλαδή της μεταβλητής Z. Έτσι οι τιμές που λείπουν μπορούν να εκτιμηθούν παρατηρώντας τις μεταβλητές X. Σε αντίστοιχο παράδειγμα με το προηγούμενο η πιθανότητα να λείπουν δεδομένα από τον αισθητήρα S, εξαρτάται από μετρήσεις άλλων μεταβλητών μετρημένων από άλλους αισθητήρες.

Όταν τα δεδομένα που λείπουν δεν κατατάσσονται σε μία από τις παραπάνω κατηγορίες ανήκουν στην MNAR. Σε αυτή την κατηγορία τα δεδομένα που λείπουν εξαρτώνται από τις ίδιες τις μεταβλητές των ελλειπόντων δεδομένων, στην προκειμένη περίπτωση βάσει της αρχικής υπόθεσης εξαρτώνται από την μεταβλητή Z. Τα δεδομένα της μεταβλητής του αισθητήρα S που αναφέρεται ως παράδειγμα έχουν ελλείπουσες τιμές της κατηγορίας MNAR όταν η τιμή μέτρησης είναι μεγαλύτερη ή μικρότερη από τα όρια μέτρησης του S.

Για την αντιμετώπιση λοιπόν των ελλειπόντων δεδομένων έχουν αναπτυχθεί αρκετές μέθοδοι οι οποίες μπορούν να κατηγοριοποιηθούν στις εξής κατηγορίες:

- Μέθοδοι διαγραφής (Deletion Methods)
- Μέθοδοι που βασίζονται σε μοντέλο (Model-based Methods)
- Μέθοδοι καταλογισμού (Imputation Methods)

Στην κατηγορία Deletion Methods ανήκει η μέθοδος Διαγραφή κατά Λίστα ή Ανάλυση Πλήρους Κατάστασης (Listwise Deletion ή Complete Case Analysis) [7, 51]. Στην συγκεκριμένη μέθοδο διαγράφονται όλες οι περιπτώσεις όπου μία τουλάχιστον τιμή των μεταβλητών που αναλύονται λείπει. Κύρια πλεονεκτήματα της μεθόδου είναι η απλότητά της καθώς και όταν τα δεδομένα που λείπουν είναι της κατηγορίας MCAR τότε η μέθοδος δεν διαστρεβλώνει τις παραμέτρους της εκτίμησης όπως τον μέσο όρο και την διακύμανση. Αντίθετα στα μειονεκτήματα της μεθόδου ανήκει η κακή εκτίμηση όταν τα δεδομένα που λείπουν δεν είναι MCAR. Ένα ακόμη μειονέκτημα είναι το γεγονός ότι δεν χρησιμοποιούνται όλα τα δεδομένα με αποτέλεσμα να μειώνεται η στατιστική ισχύς καθώς μειώνεται το πλήθος των δεδομένων.

Στην ίδια κατηγορία ανήκει και η μέθοδος Διαγραφή κατά ζεύγος ή αλλιώς γνωστή και ως Ανάλυση Διαθέσιμης Κατάστασης (Pairwise Deletion ή Available Case Analysis) [7, 51], η οποία προσπαθεί να μειώσει την απώλεια δεδομένων που εμφανίζει η προηγούμενη μέθοδος. Πιο συγκεκριμένα αναλύει όλες τις περιπτώσεις στις οποίες υπάρχουν μεταβλητές που παρουσιάζουν ενδιαφέρον. Δηλαδή στην ουσία χρησιμοποιεί κάθε διαθέσιμη πληροφορία των μεταβλητών υπολογίζοντας την μέση

τιμή κάθε μεταβλητής από τις διαθέσιμες-παρατηρούμενες τιμές της και η συνδιακύμανση υπολογίζεται από ζεύγη μεταβλητών όπου δεν παρατηρείται να λείπουν τιμές. Το γεγονός ότι η συγκεκριμένη μέθοδος χρησιμοποιεί όσο το δυνατόν περισσότερα δεδομένα, αποτελεί το μεγαλύτερο πλεονέκτημά της. Επιπλέον όπως και η Listwise Deletion μέθοδος, είναι αποδοτική όταν τα δεδομένα που λείπουν είναι της κατηγορίας MCAR. Το πιο κοινό πρόβλημα ωστόσο της συγκεκριμένης μεθόδου είναι το γεγονός ότι δεν παράγει σταθερές εκτιμήσεις καθώς ο υπολογισμός των παραμέτρων στηρίζεται στα δεδομένα που λείπουν και άρα διαφορετικά δείγματα δεδομένων παράγουν διαφορετικές εκτιμήσεις καθώς σε κάθε δείγμα τα δεδομένα που λείπουν είναι διαφορετικά.

Στην δεύτερη κατηγορία των Model-based Methods ανήκουν δύο ευρέως διαδεδομένες μέθοδοι, η μέθοδος Πολλαπλού Καταλογισμού (Multiple Imputation) και η μέθοδος Μέγιστης Πιθανοφάνειας (Maximum Likelihood). Στη Multiple Imputation (MI) μέθοδο [6, 46, 51] σε πρώτο στάδιο συμπληρώνονται δεδομένα που λείπουν χρησιμοποιώντας τιμές που προκύπτουν από ένα μοντέλο που ενσωματώνει τυχαία μεταβολή. Το συγκεκριμένο βήμα επαναλαμβάνεται m φορές με αποτέλεσμα να δημιουργούνται m σύνολα δεδομένων τα οποία είναι πανομοιότυπα όσο αφορά τα αρχικά δεδομένα αλλά διαφέρουν στα συμπληρωμένα δεδομένα που αρχικά έλλειπαν. Σε επόμενο βήμα εκτελείται η κατάλληλη ανάλυση σε κάθε σύνολο δεδομένων χωριστά. Αυτό ουσιαστικά γίνεται εφαρμόζοντας μεθόδους που θα είχαν εφαρμοστεί αν τα δεδομένα αρχικά δεν περιείχαν ελλείποντα δεδομένα [6]. Στο τελευταίο στάδιο της μεθόδου, τα αποτελέσματα του προηγούμενου βήματος, συγκεντρώνονται σε μία εκτίμηση όπως παρουσιάζεται στο [6]. Πλεονέκτημα αυτής της μεθόδου είναι ότι η διακύμανση των τιμών που λείπουν είναι πιο ακριβής κάτι στο οποίο συμβάλλει η πολλαπλή εκτίμηση, του πρώτου βήματος της μεθόδου, για κάθε τιμή που λείπει. Ένα ακόμη πλεονέκτημα, από πλευράς καλύτερης κατανόησης, είναι ο διαχωρισμός που γίνεται μεταξύ του προβλήματος των ελλειπόντων δεδομένων αρχικά και του προβλήματος των ολοκληρωμένων δεδομένων έπειτα. Το κύριο μειονέκτημα της MI μεθόδου είναι ότι αυξάνεται η πιθανότητα λαθεμένης εκτίμησης λόγω των πολλών υποθέσεων κατά την επιλογή μοντέλου.

Στην Maximum Likelihood (ML) μέθοδο σκοπός είναι ο προσδιορισμός του συνόλου των τιμών των παραμέτρων που παράγουν την υψηλότερη πιθανοφάνεια. Με άλλα λόγια η ML μέθοδος προσδιορίζει την τιμή που είναι πιο πιθανό να έχει ως αποτέλεσμα τα αρχικά παρατηρηθέντα δεδομένα. Πιο αναλυτικά [7, 8, 46], ως πρώτο βήμα της ML μεθόδου είναι η δημιουργία της συνάρτησης πιθανοφάνειας δεδομένου ότι υπάρχουν μεταβλητές χωρίς δεδομένα να λείπουν. Έτσι για να προκύψουν εκτιμήσεις μέγιστης πιθανοφάνειας αρκεί να υπολογιστούν οι παράμετροι που μεγιστοποιούν τη συνάρτηση πιθανοφάνειας. Για να αντιμετωπιστεί το πρόβλημα των ελλειπόντων δεδομένων η ML μέθοδος ακολουθεί όμοια διαδικασία. Δεδομένου ότι υπάρχουν μεταβλητές με δεδομένα που λείπουν τύπου MAR καταλήγει στην δημιουργία μίας συνάρτησης πιθανοφάνειας που περιέχει δεδομένα με τιμές που λείπουν αλλά και ολοκληρωμένα δεδομένα χωρίς να λείπουν τιμές. Στα συν της ML μεθόδου είναι το γεγονός ότι χρησιμοποιεί όλη την διαθέσιμη πληροφορία, δηλαδή και ολοκληρωμένα δεδομένα και ελλειπή δεδομένα, για τον υπολογισμό της μέγιστης πιθανοφάνειας. Ακόμη αποδίδει καλά για ελλείποντα δεδομένα των κατηγοριών MAR και MCAR. Μειονέκτημα της μεθόδου μπορεί να θεωρηθεί η δυσκολία εύρεσης του κατάλληλου μοντέλου για τον υπολογισμό των παραμέτρων που μεγιστοποιούν την συνάρτηση πιθανοφάνειας, με αποτέλεσμα την ευαισθησία των αποτελεσμάτων ως προς την ορθότητά τους ανάλογα με το μοντέλο που επιλέγεται.

Η τελευταία κατηγορία είναι αυτή των Imputation Methods και περιλαμβάνει όλες εκείνες τις

μεθόδους στις οποίες κάποια εικασία ή εκτίμηση αντικαθιστά κάθε τιμή που λείπει. Μια πολύ γρήγορη και απλή μέθοδος αυτής της κατηγορίας είναι η Υποκατάσταση Μέση Τιμής (Mean Substitution) [51], η οποία υποκαθιστά σε μία μεταβλητή, κάθε τιμή που λείπει με τον μέσο όρο των τιμών της μεταβλητής. Η συγκεκριμένη μέθοδος ωστόσο παρουσιάζει δύο σημαντικά μειονεκτήματα. Μειώνει την μεταβλητότητα του δείγματος της μεταβλητής καθώς αντικαθιστά την ίδια τιμή σε όλες τις τιμές που λείπουν και επιπλέον αποδυναμώνει τις εκτιμήσεις της συνδιακύμανσης και της συσχέτισης των δεδομένων γιατί αγνοεί την σχέση μεταξύ των μεταβλητών.

Μία ακόμη μέθοδος που ανήκει στην ίδια κατηγορία είναι η μέθοδος του Καταλογισμού με Παλινδρόμηση (Regression Imputation), η οποία αντικαθιστά τις τιμές που λείπουν με το αποτέλεσμα της πρόβλεψης που προκύπτει από μία εξίσωση παλινδρόμησης. Πιο συγκεκριμένα [21, 44, 51], η παλινδρόμηση (regression) είναι μία στατιστική τεχνική μοντελοποίησης που χρησιμοποιείται ευρέως με σκοπό να περιγράψει την σχέση μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσοτέρων ανεξάρτητων μεταβλητών. Στηρίζεται στο γεγονός ότι τα δεδομένα που εξετάζονται ταιριάζουν με μερικά γνωστά είδη συναρτήσεων, έτσι ως πρώτο βήμα η συγκεκριμένη μέθοδος καθορίζει την καλύτερη συνάρτηση που είναι ικανή να μοντελοποιήσει τα δεδομένα που έχουν δοθεί. Οι προβλέψεις για τις ελλιπείς περιπτώσεις υπολογίζονται σύμφωνα με το προσαρμοσμένο μοντέλο και χρησιμεύουν ως αντικαταστάσεις των ελλειπόντων δεδομένων. Το κύριο πλεονέκτημα του Regression Imputation είναι ότι χρησιμοποιούνται πληροφορίες που παρέχονται από τα παρατηρούμενα δεδομένα ώστε να δημιουργηθεί το μοντέλο. Ένα μειονέκτημα αυτής της μεθόδου είναι ότι μπορεί να προκύψει υπερεκτίμηση του μοντέλου προκαλώντας υπερπροσαρμογή στα δεδομένα εκπαίδευσης (overfitting) με αποτέλεσμα να χάνεται η γενίκευση που μπορεί να παρέχει το μοντέλο ώστε να περιγράψει ασφαλέστερα το σύνολο των δεδομένων.

2.2 Πρόβλεψη Πληρότητας Κτηρίου

Η ικανότητα πρόβλεψης του πότε σε ένα κτήριο υπάρχει παρουσία ατόμων και πότε όχι αποτελεί βασική απαίτηση για το σωστό έλεγχο των ενεργειακών συστημάτων του. Η πρόβλεψη αυτή συμβάλει στον έλεγχο του αισθήματος άνεσης στο χώρο αλλά και στη δραστική μείωση της καταναλισκόμενης ενέργειας. Για παράδειγμα ένα σύστημα θέρμανσης μπορεί να ενεργοποιηθεί ένα μικρό χρονικό διάστημα πριν από την εκτιμώμενη, από την πρόβλεψη, ώρα άφιξης των ενοίκων και ενώ ήταν απενεργοποιημένο για μεγάλο χρονικό διάστημα, αποφεύγοντας την έλλειψη άνεσης. Επιπλέον με την ενέργεια αυτή, και όχι προθερμαίνοντας το χώρο για μεγάλη χρονική περίοδο προσπαθώντας να διατηρηθεί η σωστή θερμοκρασία, επιτυγχάνεται μείωση της ενέργειας που καταναλώνεται.

Έτσι λοιπόν για τον έλεγχο των ενεργειακών συστημάτων κτηρίων έχουν αναπτυχθεί διάφορες μέθοδοι πρόβλεψης πληρότητας κτηρίων όπως στις εργασίες [4, 19, 24, 28, 30, 33, 48, 50]. Ανάλογα με την λογική που ακολουθούν μπορεί να γίνει ένας διαχωρισμός μεταξύ των μεθόδων. Μπορούν να χωριστούν σε μεθόδους που βασίζονται στην τρέχουσα κατάσταση των ενοίκων και σε μεθόδους που βασίζονται σε παλαιότερα χρονοδιαγράμματα πληρότητας του κτηρίου.

Η Κατηγορία των μεθόδων που βασίζονται στην τρέχουσα κατάσταση των ενοίκων ονομάζεται επίγνωσης πλαισίου (context-aware) καθώς οι μέθοδοι αυτής της κατηγορίας στηρίζονται στο τρέχον πλαίσιο κατάστασης των ενοίκων δηλαδή στηρίζονται στην τοποθεσία που βρίσκονται οι ένοικοι ή με το τι δραστηριότητα απασχολούνται. Για παράδειγμα ο αλγόριθμος που παρουσιάζεται στην εργασία

[28] εκτιμά το χρόνο άφιξης των ενοίκων στο κτήριο σύμφωνα με την τρέχουσα θέση και την πορεία που ακολουθούν για να επιστρέψουν. Η θέση τους ορίζεται χρησιμοποιώντας συστήματα GPS που είναι ενσωματωμένα σε ειδικές συσκευές ή στα κινητά τηλέφωνα των ενοίκων. Έπειτα χρησιμοποιούνται υπηρεσίες χαρτογράφησης από το διαδίκτυο με σκοπό να καθορίζεται η απόσταση που απέχουν από το κτήριο και ο εκτιμώμενος χρόνος μέχρι να φτάσουν σε αυτό. Παρόμοια μεθοδολογία ακολουθείται και σε άλλες εργασίες όπως η [45] και η [54] οι οποίες εστιάζουν κυρίως στο πότε για πόσο και ποιο μέρος θα επισκεφτεί ένας άνθρωπος παρά στην πληροφορία της πληρότητας ενός κτηρίου. Σκοπός των εργασιών αυτών είναι εκμεταλλευόμενες την δυνατότητα πρόβλεψης των μελλοντικών θέσεων των ανθρώπων, να τους παρέχουν συγκεκριμένες πληροφορίες, διαφημίσεις και γενικότερα αναφορές σχετικές με το μέρος που πρόκειται να επισκεφτούν.

Από την άλλη μεριά, η άλλη κατηγορία των μεθόδων πρόβλεψης πληρότητας, λόγω του ότι στηρίζεται σε παλαιότερα χρονοδιαγράμματα πληρότητας κτηρίων ονομάζεται, μέθοδοι βασισμένες σε χρονοδιαγράμματα (schedule-based methods). Στην βιβλιογραφία έχουν αναπτυχθεί μέθοδοι όπως στις εργασίες [30, 33, 48], που βασίζονται μόνο σε δεδομένα πληρότητας ενός κτηρίου τα οποία έχουν συλλεχθεί σε μία παλαιότερη εκτεταμένη χρονική περίοδο. Μια μέθοδος που ανήκει σε αυτή την κατηγορία είναι αυτή της εργασίας [30] όπου παρουσιάζεται ο αλγόριθμος εκτίμησης των Πιθανοτήτων Παρουσίας (Presence Probabilities). Στην προκειμένη περίπτωση η πληρότητα ενός κτηρίου ανιχνεύεται χρησιμοποιώντας συσκευές GPS που έχουν μαζί τους οι ένοικοι. Το κτήριο θεωρείται ότι κατοικείται όταν η συσκευή δείχνει ότι ο ένοικος απέχει λιγότερο από 100 μέτρα από το κτήριο. Χρησιμοποιώντας τα δεδομένα του GPS η συγκεκριμένη μέθοδος υπολογίζει την πιθανότητα (p_{away}) το κτήριο να είναι ακατοίκητο οποιαδήποτε χρονική στογμή της ημέρας μίας εβδομάδας. Οι τιμές της p_{away} καταχωρούνται σε ένα διάνυσμα το οποίο αντιπροσωπεύει την πιθανότητα να μην κατοικείται το κτήριο κατα μήκος μιας εβδομάδας. Η πιθανότητα σε κάθε θέση του διανύσματος εξομαλύνεται χρησιμοποιώντας τιμές των προηγούμενων και μετέπειτα θέσεων μέσω μιας σταθεράς εξομάλυνσης s και ενός παράγοντα κανονικοποίησης λ_s.

Μία άλλη μέθοδος της ίδιας κατηγορίας είναι η μέθοδος Έξυπνου Θερμοστάτη (Smart Thermostat) της εργασίας [33]. Σε αυτή τη μέθοδο η κατάσταση πληρότητας ενός κτηρίου καθορίζεται με την χρήση Κρυμμένου Μαρκοβιανού μοντέλου (Hidden Markov Model) που περιγράφεται αναλυτικά στην Παράγραφο 3.10.1. Το μοντέλο αυτό επιτρέπει την εκτίμηση του πότε ένα κτήριο είναι κατοικημένο ή ακατοίκητο. Για να πραγματοποιήσει την εκτίμηση δέχεται ως είσοδο δεδομένα από παλαιότερα χρονοδιαγράμματα πληρότητας όπως και πραγματικά δεδομένα που συλλέγονται από αισθητήρες μέσα στο κτήριο. Όταν το κτήριο εκτιμάται ως ακατοίκητο ο αλγόριθμος απενεργοποιεί για παράδειγμα το σύστημα θέρμανσης. Προβλέποντας πότε θα κατοικηθεί εκ νέου το κτήριο ο αλγόριθμος το προθερμαίνει για ένα χρονικό διάστημα πριν την προβλεπόμενη ώρα της επανακατοίκησης του από τους ενοίκους, πραγματοποιώντας έτσι μείωση του κινδύνου απώλειας άνεσης των ενοίκων, οι οποίοι σε διαφορετική περίπτωση όπου δεν θα χρησιμοποιούνταν η συγκεκριμένη μέθοδος θα διέμεναν σε ένα ψυχρό χώρο έως ότου αυτός ζεσταθεί.

Στην εργασία [48], ο αλγόριθμος Προθέρμανσης (Preheat) είναι ένα ακόμη παράδειγμα μεθόδου που βασίζεται σε παλαιότερα χρονοδιαγράμματα πληρότητας ενός κτηρίου. Ο συγκεκριμένος αλγόριθμος διατηρεί ένα διάνυσμα για την αποθήκευση της πραγματικής κατάστασης πληρότητας του κτηρίου για την τρέχουσα ημέρα αρχίζοντας από τα μεσάνυχτα. Για να προβλέψει την πληρότητα του κτηρίου από μια δεδομένη χρονική στιγμή της τρέχουσας ημέρας και έπειτα, ο αλγόριθμος υπολογίζει

την απόσταση Hamming μεταξύ της κατάστασης πληρότητας του κτηρίου μέχρι τη δεδομένη χρονική στιγμή της τρέχουσας ημέρας και τα αντίστοιχα τμήματα των διανυσμάτων πληρότητας του κτηρίου που αφορούν δεδομένα παλαιότερων ημερών.

Στην παρούσα διπλωματική εργασία μελετήθηκαν δύο μέθοδοι πρόβλεψης πληρότητας ενός κτηρίου που βασίστηκαν στις εργασίες [48] και [33]. Χρησιμοποιήθηκαν αυτές οι εργασίες καθώς η λογική τους στηρίζεται αποκλειστικά σε παλαιότερα δεδομένα. Έτσι ήταν εφικτό να αναπτυχθούν και να αξιολογηθούν αυτές οι μέθοδοι λόγω της δυνατότητας χρήσης πραγματικών δεδομένων κτιρίων που έχουν καταγραφεί για μεγάλα χρονικά διαστήματα στο παρελθόν. Δεν χρησιμοποιήθηκαν οι υπόλοιπες μέθοδοι διότι απαιτούσαν επιπλέον εξοπλισμό, όπως συστήματα GPS, που δεν ήταν διαθέσιμος.

Κεφάλαιο 3

Μέθοδοι Διόρθωσης Δεδομένων και Πρόβλεψης Πληρότητας Κτηρίου

Στο κεφάλαιο αυτό περιγράφονται διάφοροι τρόποι διόρθωσης δεδομένων (data correction) και πιο συγκεκριμένα διόρθωση ελλειπόντων δεδομένων(missing data) όπως επίσης και μέθοδοι πρόβλεψης πληρότητας (occupancy prediction) ενός κτηρίου που στην ουσία προβλέπουν τις χρονικές περιόδους στις οποίες κατοικείται ένα κτήριο άρα υπάρχει και χρήση των συστημάτων του.

3.1 Γενική Περιγραφή της Διόρθωσης Δεδομένων

Σε αυτή την ενότητα παρουσιάζονται αναλυτικά μέθοδοι παλινδρόμησης (regression) και προσαρμογής καμπύλης (curve fitting) που χρησιμοποιήθηκαν στην παρούσα διπλωματική ως μέθοδοι διόρθωσης δεδομένων. Σκοπός της χρήσης των μεθόδων για διόρθωση δεδομένων είναι να προβλεφθούν τιμές που μπορεί να λείπουν από ένα σύνολο τιμών που επιστρέφουν διάφοροι αισθητήρες προσαρμοσμένοι σε ένα κτήριο. Οι τιμές αυτές μπορεί να λείψουν λόγω αστοχίας κάποιου αισθητήρα, προσωρινή βλάβη, διακοπή της ηλεκτροδότησης του κτηρίου για ένα χρονικό διάστημα μέσα στην μέρα ή για οποιοδήποτε άλλο λόγο. Με αυτό τον τρόπο πετυχαίνεται μια ολοκληρωμένη ροή δεδομένων και συνεχής ενημέρωση που περιγράφει την κατάσταση ενός κτηρίου με αποτέλεσμα την επίτευξη βέλτιστου ελέγχου και προσαρμογή των συστημάτων του κτηρίου άρα και αποδοτικότερη αξιοποίηση τους.

Η regression μέθοδος είναι μία στατιστική τεχνική μοντελοποίησης που χρησιμοποιείται ευρέως με σκοπό να περιγράψει την σχέση μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσοτέρων ανεξάρτητων μεταβλητών. Στα μοντέλα παλινδρόμησης δεδομένου ότι υπάρχει ένα διάνυσμα ανεξάρτητων μεταβλητών X, ένα διάνυσμα εξαρτημένων μεταβλητών Y και δηλώνοντας ως β τις παραμέτρους συσχέτισης, επιδιώκεται η συσχέτιση του Y ως συνάρτηση των X και β δηλαδή $Y \simeq F(X, \beta)$. Ο συνήθης φορμαλισμός είναι $E(Y|X) = f(X, \beta)$. Οι μέθοδοι της παλινδρόμησης που χρησιμοποιήθηκαν στην παρούσα διπλωματική είναι:

1. Πολυωνυμική παλινδρόμηση (Polynomial Regression)
2. Τοπική παλινδρόμηση (Local Regression)
3. Παλινδρόμηση με χρήση Support Vector (Support Vector Regression)

4. Gaussian processes

Παρόμοια με το regression οι τεχνικές για curve fitting είναι διαδικασίες που κατασκευάζουν μία καμπύλη ή μία μαθηματική συνάρτηση με σκοπό την όσο δυνατόν καλύτερη προσαρμογή σε μια σειρά από σημεία δεδομένων. Οι προσαρμοσμένες καμπύλες μπορούν να χρησιμοποιηθούν για την οπτικοποίηση των δεδομένων, να εξάγουν τιμές μιας συνάρτησης όπου δεν υπάρχουν διαθέσιμα δεδομένα και να συνοψίσει τις σχέσεις μεταξύ δύο ή περισσοτέρων μεταβλητών. Στην παρούσα εργασία χρησιμοποιήθηκαν οι εξής τρόποι:

1. Προσαρμογή με Νευρωνικό Δίκτυο (Neural Network fitting)
2. Καταλογισμός πλησιέστερου γείτονα (Nearest Neighbor imputation)

3.2 Πολυωνυμική Παλινδρόμηση (Polynomial regression)

Το Polynomial regression [44] είναι μία μορφή γραμμικής παλινδρόμησης, δηλαδή μία προσέγγιση για την αναπαράσταση της σχέσης μίας εξαρτημένης βαθμωτής μεταβλητής και μίας ή περισσοτέρων ανεξάρτητων μεταβλητών. Πιο συγκεκριμένα, στο Polynomial regression η σχέση μεταξύ των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής διαμορφώνεται ως ένα πολυώνυμο. Το γενικό μοντέλο πολυωνυμικής παλινδρόμησης είναι το εξής:

$$y = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \ldots + \alpha_m x_i^n \ , \ i = 1, 2, \ldots, n \tag{3.1}$$

Για μεγαλύτερη ευκολία, τα μοντέλα αυτά θεωρούνται ότι είναι όλα γραμμικά από την άποψη της εκτίμησης, αφού η συνάρτηση παλινδρόμησης είναι γραμμική σε σχέση με τις άγνωστες παραμέτρους $\alpha_0, \alpha_1, \ldots, \alpha_m$. Ως εκ τούτου, για την ανάλυση ελαχίστων τετραγώνων, τα υπολογιστικά και επαγωγικά προβλήματα του polynomial regression μπορούν να αντιμετωπιστούν χρησιμοποιώντας τις τεχνικές της πολλαπλής παλινδρόμησης. Αυτό συμβαίνει με μεταχείριση των $x, x^2, \ldots, x^m$ ως διακριτών ανεξάρτητων μεταβλητών σε ένα μοντέλο πολλαπλής παλινδρόμησης. (Πολλαπλή παλινδρόμηση συμβαίνει όταν υπάρχουν περισσότερες από μία ανεξάρτητες μεταβλητές.)

Έτσι μπορούμε να εκφράσουμε το polynomial regression μοντέλο 3.1 σε μορφή πίνακα, δηλαδή σχεδιασμό ενός πίνακα X, ενός διανύσματος απόκρισης y και ενός διανύσματος παραμέτρου α. Η i-οστή σειρά των X και y περιέχουν την τιμή των x και y για το i-οστό δείγμα δεδομένων. Οπότε το μοντέλο μπορεί να γραφτεί ως εξής:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ 1 & x_3 & x_3^2 & \ldots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix}$$

Πιο απλά μπορεί να γραφτεί ως: $y = X\alpha$. Το διάνυσμα των εκτιμώμενων πολυωνυμικών παραμέτρων παλινδρόμησης (χρησιμοποιώντας συνήθη εκτίμηση ελαχίστων τετραγώνων) είναι:

$$\hat{\alpha} = (X^T X)^{-1} X^T y \tag{3.2}$$

3.3 Τοπική Παλινδρόμηση (Local regression)

Το Local regression είναι μια μέθοδος μη παραμετρικής παλινδρόμησης, δηλαδή, μέθοδος στην οποία η δομή του μοντέλου δεν έχει καθοριστεί. Αυτού του είδους η παλινδρόμηση αποδίδει γύρω από ένα σημείο ενδιαφέροντος, χρησιμοποιώντας ως δεδομένα εκπαίδευσης του μοντέλου μόνο εκείνα τα δεδομένα που είναι τοπικά στο σημείο ενδιαφέροντος. Η μέθοδος αυτή στην ουσία εκτελεί μια "τοπική" γραμμική παλινδρόμηση επί των σημείων αυτών στο σύνολο των δεδομένων. Τα δεδομένα που χρησιμοποιούνται ως δεδομένα εκπαίδευσης, σταθμίζονται από έναν πυρήνα (kernel). Αυτός είναι και ο λόγος που το local regression λέγεται και Kernel regression.

Στην παρούσα εργασία χρησιμοποιήθηκε ένας αλγόριθμος [3] που προσέγγισή του είναι να κατασκευάσει ένα τοπικό μέσο εκτιμητή (local mean estimator) για τα δεδομένα προσαρμόζοντας σε αυτά τοπική γραμμική παλινδρόμηση (local linear regression). Το σύνηθες Local regression είναι στην ουσία ένας τοπικός σταθερός εκτιμητής. Η επέκταση του τοπικού εκτιμητή λαμβάνεται με την επίλυση του προβλήματος των ελαχίστων τετραγώνων ως εξής:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} y - \alpha - \beta(x_i - x)^2 w(x_i - x; h) \tag{3.3}$$

όπου, *h* είναι η παράμετρος εξομάλυνσης που ελέγχει το πλάτος του πυρήνα της συνάρτησης *w* και *w* είναι η Gaussian συνάρτηση.

Ο local mean estimator μπορεί να δοθεί με την διατύπωση του παρακάτω τύπου:

$$\hat{m(x)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\{s_2(x;h) - s_1(x;h)(x_i - x)\} w(x_i - x; h) y_i}{s_2(x;h)s_0(x;h) - s_1(x;h)^2} \tag{3.4}$$

όπου $s_r(x;h) = \sum\{(x_i - x)^r w(x_i - x; h)\}/n$.

Η Gaussian συνάρτηση *w* είναι της μορφής:

$$w(x) = ae^{-\frac{(x-b)^2}{2c^2}} + d \quad a, b, c, d \in \mathbb{R} \tag{3.5}$$

Το ολοκλήρωμα της Gaussian συνάρτησης είναι η συνάρτηση σφάλματος. Χρησιμοποιώντας το ολοκλήρωμα της Gaussian συνάρτησης

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}, \tag{3.6}$$

προκύπτει

$$\int_{-\infty}^{\infty} ae^{-\frac{(x-b)^2}{2c^2}} dx = ac \cdot \sqrt{2\pi} \tag{3.7}$$

Το ολοκλήρωμα γίνεται 1 μόνο όταν $a = 1/c\sqrt{2\pi}$, και σε αυτή την περίπτωση η Gaussian είναι η συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής τυχαίας μεταβλητής με αναμενόμενη τιμή $\mu = b$ και διακύμανση $\sigma^2 = c^2$:

$$w(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{3.8}$$

Προκειμένου να κατασκευαστεί μια εκτίμηση πυκνότητας από παρατηρηθέντα δεδομένα, είναι αναγκαίο να επιλεχθεί η βέλτιστη τιμή της παραμέτρου εξομάλυνσης h. Ένα συνολικό μέτρο της αποτελεσματικότητας της $\hat{f}$ στην εκτίμηση μιας συνάρτησης f παρέχεται από το Μέσο Ολοκληρωμένο Τετραγωνικό Σφάλμα (Mean Integrated Squared Error, MISE) που περιγράφεται από το βιβλίο [17] και δίνεται από τον τύπο:

$$MISE(\hat{f}) = E\{\int [\hat{f}(y) - f(y)]^2 dy\} \tag{3.9}$$

Ειδικότερα με διάφορες προσεγγίσεις που γίνονται στο βιβλίο [17] το MISE μπορεί να γραφτεί ως εξής:

$$MISE(\hat{f}) \approx \frac{1}{4} h^4 \sigma_w^4 \int f''(y)^2 dy + \frac{1}{nh}\alpha(w). \tag{3.10}$$

Απο την παραπάνω κατά προσέγγιση έκφραση για το MISE προκύπτει ότι η τιμή του h η οποία ελαχιστοποιεί το MISE σε μία ασυμπτωτική έννοια είναι:

$$h_{opt} = \{\frac{\gamma(w)}{\beta(f)n}\}^{\frac{1}{5}} \tag{3.11}$$

όπου, $\gamma(w) = \alpha(w)/\sigma_w^4$ και $\beta(f) = \int f''(y)^2 dy$. Ο παραπάνω τύπος για τον υπολογισμό της βέλτιστης τιμής του h δεν μπορεί να χρησιμοποιηθεί άμεσα στην πράξη καθώς περιλαμβάνει την άγνωστη συνάρτηση πυκνότητας. Ωστόσο, είναι πολύ κατατοπιστικός δείχνοντας πώς η παράμετρος εξομάλυνσης h, μειώνεται με το μέγεθος του δείγματος n, αναλογικά ως προς $n^{-\frac{1}{5}}$. Επίσης, παρουσιάζει και την επίδραση της καμπυλότητας της f διαμέσου του παράγοντα $\beta(f)n$.

Ο υπολογισμός της βέλτιστης τιμής του h [17] όταν μια συνάρτηση f έχει κανονική κατανομή πυκνότητας γίνεται ως εξής:

$$h = (\frac{4}{3n})^{1/5}\sigma \tag{3.12}$$

με το σ να δηλώνει την τυπική απόκλιση της κατανομής. Για να επιτευχθεί όμως ο υπολογισμός του βέλτιστου h και σε περιπτώσεις όπου δεν έχουμε κανονική κατανομή πυκνότητας είναι προτιμότερη η εκτίμηση του σ χρησιμοποιώντας την μέθοδο του εκτιμητή μέσης απόλυτης απόκλισης (median absolute deviation estimator)

$$\tilde{\sigma} = \text{median}\{|y_i - \tilde{\mu}|\}/0.6745, \tag{3.13}$$

όπου $\tilde{\mu}$ δηλώνει το διάμεσο(median) του δείγματος.

Ολοκληρώνοντας όσον αφορά την περιγραφή του Local regression αξίζει να σημειωθεί πως η παραπάνω προσέγγιση του local linear estimator μπορεί να βελτιώσει την εκτίμηση κοντά στην άκρη της περιοχής επί της οποίας έχουν συλλεχθεί τα δεδομένα.

3.4 Παλινδρόμηση Διανύσματος Υποστήριξης (Support Vector Regression)

Το Support Vector Regression (SVR) [25, 49] είναι μια μέθοδος που βασίζεται στις μηχανές διανύσματος υποστήριξης (Support Vector Machines (SVM)) τα οποία είναι μοντέλα που συνδέονται με αλγορίθμους μηχανικής μάθησης που αναλύουν δεδομένα. Ένα κύριο χαρακτηριστικό της SVR μεθόδου είναι ότι αντί της ελαχιστοποίησης του σφάλματος της εκπαίδευσης που παρατηρήθηκε,

η SVR μέθοδος προσπαθεί να ελαχιστοποιήσει ένα γενικευμένο όριο λάθους έτσι ώστε να επιτευχθεί μία πιο γενικευμένη απόδοση. Αυτό το γενικευμένο όριο σφάλματος είναι ο συνδυασμός του σφάλματος εκπαίδευσης και ενός όρου κανονικοποίησης ο οποίος ελέγχει την πολυπλοκότητα του χώρου υπόθεσης. Σκοπός δηλαδή της SVR μεθόδου, δεδομένου ενός συνόλου δεδομένων εκπαίδευσης $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset X \times \mathbb{R}$ (με X να δηλώνει το χώρο των δεδομένων εισόδου), είναι να βρεθεί μία συνάρτηση $f(x)$ η οποία να έχει το πολύ ε απόκλιση από τους παρατηρούμενους στόχους y_i για όλα τα δεδομένα εκπαίδευσης και την ίδια στιγμή να είναι όσο το δυνατόν πιο επίπεδη.

Πιο αναλυτικά, στην SVR μέθοδο η είσοδος x αρχικά χαρτογραφείται επάνω σε ένα m-διάστατο χώρο χαρακτηριστικών χρησιμοποιώντας μια σταθερή μη γραμμική χαρτογράφηση και στην συνέχεια ένα γραμμικό μοντέλο κατασκευάζεται σε αυτό το χώρο. Το γραμμικό μοντέλο περιγράφεται ως εξής:

$$f(x) = \langle w, x \rangle + b \tag{3.14}$$

με $w \in X$ και $b \in \mathbb{R}$ όπου $\langle \cdot, \cdot \rangle$ δηλώνει το εσωτερικό γινόμενο στο X. Η μείωση της πολυπλοκότητας του μοντέλου, δηλαδή να γίνει όσο το δυνατόν πιο ομαλό, πραγματοποιείται ελαχιστοποιώντας το $\|w\|^2 = \langle w, w \rangle$. Έτσι προκύπτει το εξής πρόβλημα ελαχιστοποίησης:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 \\ &\text{subject to} \quad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \end{aligned} \tag{3.15}$$

Αυτό που συμβαίνει στην (3.15) είναι ότι προσεγγίζει τα ζεύγη (x_i, y_i) με ακρίβεια ε. Επειδή όμως υπάρχει χρήσιμη πληροφορία και εκτός του ε, χρησιμοποιούνται μη αρνητικές μεταβλητές ξ_i, ξ_i^* με $i = 1, 2, \ldots, n$ για να μετρηθεί η απόκλιση των δειγμάτων εκπαίδευσης εκτός της ζώνη ε που δημιουργείται εκατέρωθεν των (x_i, y_i). Πιο συγκεκριμένα το ξ_i περιγράφει τα δείγματα πάνω από την ζώνη και το ξ_i^* περιγράφει τα δείγματα κάτω από την ζώνη. Έτσι το πρόβλημα ελαχιστοποίησης της (3.15) διαμορφώνεται ως εξής:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*), \\ &\text{subject to} \quad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \ , \ i = 1, \ldots, n \end{cases} \end{aligned} \tag{3.16}$$

όπου $C > 0$ είναι μία σταθερά που καθορίζει την σχέση μεταξύ της ομαλότητας της f και του ποσοστού μέχρι το οποίο είναι αποδεκτές αποκλίσεις μεγαλύτερες από ε.

Πρακτικά είναι συνήθως δύσκολο να προσδιοριστεί ένα κατάλληλο μοντέλο για τον προσδιορισμό του ε. Έτσι για το λόγο αυτό στην παρούσα διπλωματική εργασία εφαρμόζεται η μέθοδος ν-SVR [47] η οποία χρησιμοποιεί την παράμετρο $0 \leq \nu < 1$ για να επιτευχθεί ισορροπία μεταξύ της ζώνης ε, της ομαλότητας του μοντέλου και των μεταβλητών ξ, ξ^*. Έτσι το πρόβλημα ελαχιστοποίησης γίνεται:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\nu\varepsilon + C\,\frac{1}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*), \\ &\text{subject to} \quad \begin{cases} y_i - \langle w, x_i\rangle - b \le \varepsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0\ ,\ i = 1, \ldots, n \end{cases} \end{aligned} \tag{3.17}$$

Αυτό το πρόβλημα βελτιστοποίησης μπορεί να μετατραπεί σε πρόβλημα διπλής βελτιστοποίησης με την χρήση τεχνικών των πολλαπλασιαστών Lagrange [13, 14] και με εφαρμογή των συνθηκών Karush-Kuhn-Tucker [14] προκύπτει ότι:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b\ . \tag{3.18}$$

Το επόμενο βήμα είναι ο αλγόριθμος να γίνει μη γραμμικός. Παρατηρώντας την (3.18) φαίνεται ότι ο αλγόριθμος εξαρτάται από το εσωτερικό γινόμενο των x_i. Έτσι βάσει του [47] αρκεί να είναι γνωστή η συνάρτηση πυρήνα $\mathcal{K}(x_i, x_j) : \langle \Phi(x_i), \Phi(x_j)\rangle$ αντί της Φ. Οπότε το πρόβλημα βελτιστοποίησης (3.18) γίνεται:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathcal{K}(x_i, x) + b\ . \tag{3.19}$$

Στην παρούσα διπλωματική χρησιμοποιείται ως συνάρτηση πυρήνα η radial-basis function (RBF) που είναι της μορφής:

$$\mathcal{K}(x_i, x) = \exp(-\gamma\|x_i - x\|^2). \tag{3.20}$$

3.5 Gaussian Διαδικασίες (Gaussian Processes)

Έστω μία συνάρτηση πυκνότητας πιθανότητας $p(f)$ που ορίζεται σε ένα χώρο συναρτήσεων F. Παίρνοντας συναρτήσεις f από το χώρο F σύμφωνα με το $p(f)$ δημιουργούνται δείγματα μίας στοχαστικής διαδικασίας. Τα δείγματα αυτά είναι τυχαίες συναρτήσεις της κατανομής που έχει συνάρτηση πυκνότητας πιθανότητας $p(f)$. Έτσι σύμφωνα με την εργασία [18] μία διαδικασία Gaussian είναι ένα υποσύνολο του συνόλου των στοχαστικών διαδικασιών που έχουν την ιδιότητα, η από κοινού κατανομή σε κάθε πεπερασμένο σύνολο σταθερών σημείων δοκιμής $\mathbf{X}$ να είναι μία πολυπαραγοντική Gaussian. Δηλαδή η κατανομή των $f \in \mathbb{R}^n$ είναι μία πολυπαραγοντική Gaussian για όλα τα πεπερασμένα n και όλα τα $x_i \in \mathbf{X}$.

Πιο αναλυτικά [18, 41], είναι σύνηθες σε πραγματικές καταστάσεις μοντελοποίησης να μην είναι γνωστές αυτές καθ΄αυτές οι τιμές των συναρτήσεων αλλά να είναι γνωστές μαζί με θόρυβο $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ ο οποίος ακολουθεί ανεξάρτητη ομοιόμορφη κατανομή. Είναι δηλαδή:

$$y = f(x) + \varepsilon. \tag{3.21}$$

Για παλινδρόμηση με διαδικασία Gaussian χρησιμοποιείται μηδενικής μέσης τιμής πολυπαραγοντική κατανομή Gaussian :

$$\mathbf{f}|\mathbf{X}, \theta\ \sim\ \mathcal{N}(\mathbf{0}, \mathbf{K}), \tag{3.22}$$

όπου **K** είναι ένας $n \times n$ πίνακας συνδιακύμανσης που εξαρτάται από το **X** και τα hyperparameters θ. Επιπλέον ισχύει ότι :

$$\mathbf{K}_{i,j} = k(x_i, x_j), \tag{3.23}$$

όπου $k(\,\cdot\,,\,\cdot\,)$ είναι μία συνάρτηση παραμετροποιήσιμη από το θ και ονομάζεται συνάρτηση συνδιακύμανσης (covariance function).

Έτσι γνωρίζοντας ένα πλήθος δεδομένων και την συνάρτηση συνδιακύμανσης χρησιμοποιείται το μοντέλο της Gaussian Process για να κάνει πρόβλεψη τιμών. Πιο συγκεκριμένα, η από κοινού κατανομή των δεδομένων εκπαίδευσης f και των δεδομένων δοκιμής f_* χρησιμοποιώντας την (3.22) είναι:

$$\begin{bmatrix} f \\ f_* \end{bmatrix} |\mathbf{X}, \theta \sim \mathcal{N}(0, \begin{bmatrix} \mathbf{K} & \mathbf{k} \\ \mathbf{k}^T & \kappa \end{bmatrix}), \tag{3.24}$$

όπου $\mathbf{k} = [k(x_*, x_1) \cdots k(x_*, x_n)]^T$ είναι το διάνυσμα που διαμορφώθηκε απο την συνδιακύμανση μεταξύ των ζευγαριών των σημείων x_* και των δεδομένων εκπαίδευσης. Ακόμη, $\kappa = k(x_*, x_*)$.

Χρησιμοποιώντας την (3.21), η από κοινού κατανομή των παρατηρούμενων δεδομένων y και των παρατηρούμενων δεδομένων δοκιμής f_* γίνεται:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} |\mathbf{X}, \theta \sim \mathcal{N}(0, \begin{bmatrix} \mathbf{K} + \sigma^2 I & \mathbf{k} \\ \mathbf{k}^T & \kappa + \sigma^2 \end{bmatrix}), \tag{3.25}$$

Γνωρίζοντας ότι από κοινού κατανομή είναι Gaussian και σύμφωνα με το παράρτημα Α.2 του [41] προκύπτει ότι η προγνωστική μέση τιμή και διακύμανση είναι:

$$\bar{f}_* = \mathbf{k^T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \tag{3.26}$$

$$v(f_*) = \kappa + \sigma^2 - \mathbf{k^T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}. \tag{3.27}$$

Με αυτό τον τρόπο λοιπόν δίνεται η δυνατότητα υπολογισμού της Gaussian προγνωστικής κατανομής για κάθε σημείο δοκιμής x_*. Για τον υπολογισμό Gaussian προγνωστικής κατανομής ενός συνόλου m σημείων δοκιμής χρησιμοποιούνται τα εξής:

$$\bar{f}_* = \mathbf{K_*^T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\,\mathbf{y} \tag{3.28}$$

$$v(f_*) = \mathbf{K}_{**} + \sigma^2 - \mathbf{K_*^T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_*. \tag{3.29}$$

όπου $\mathbf{K}_*$ είναι ένας n x m πίνακας των διακυμάνσεων μεταξύ των εισόδων εκπαίδευσης και των σημείων δοκιμής. Ο πίνακας $\mathbf{K}_{**}$ είναι m x m και περιέχει τις συνδιακυμάνσεων μεταξύ των σημείων δοκιμής.

Η οριακή πιθανότητα (marginal likelihood) ενός μοντέλου Gaussian process είναι το ολοκλήρωμα του γινομένου της συνάρτησης πιθανοφάνειας και της αρχικής πυκνότητας πιθανότητας.

$$p(\mathbf{y}|\mathbf{X}, \theta, \sigma^2) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X}, \theta, \sigma^2) \cdot p(\mathbf{f}|\mathbf{X}, \theta)\mathrm{d}\mathbf{f} \tag{3.30}$$

Σύμφωνα με το μοντέλο της Gaussian process ισχύει ότι $\mathbf{f}|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, ή:

$$\log\, p(\mathbf{f}|\mathbf{X}, \theta, \sigma^2) = -\frac{1}{2}\mathbf{f^T}K^{-1}\mathbf{f} - \frac{1}{2}\log|K| - \frac{n}{2}\log\, 2\pi, \tag{3.31}$$

και η πιθανοφάνεια είναι $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2 I)$ οπότε χρησιμοποιώντας τα παραρτήματα Α.7 και Α.8 [41] αποδίδεται η log marginal likelihood:

$$\log p(\mathbf{y}|X, \theta, \sigma^2) = -\frac{1}{2}\mathbf{y}^T(K + \sigma^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma^2 I| - \frac{n}{2}\log 2\pi. \tag{3.32}$$

Στον αλγόριθμο που χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία [40], οι hyperparameters αποτελούν ελεύθερες παράμετροι των συναρτήσεων μέσης τιμής (mean function) και συνδιακύμανσης (covariance function) όπως επίσης και των συναρτήσεων πιθανότητας (likelihood function), οι οποίες προσδιορίζουν το Gaussian process. Όταν καθορίζονται οι hyperparameters είναι σημαντικό ο αριθμός των στοιχείων για κάθε ομάδα συναρτήσεων, που αναφέρθηκαν παραπάνω, να αντιστοιχεί στον αριθμό των παραμέτρων που απαιτούνται από τις mean, covariance και likelihood συναρτήσεις αντίστοιχα.

Η mean function $m_\Phi : X \to \mathbb{R}$ με hyperparameter Φ, μιας Gaussian Process f είναι μία βαθμωτή συνάρτηση που ορίζεται από ολόκληρο το X και υπολογίζει την αναμενόμενη τιμή

$$m(\mathbf{x}) = E[f(\mathbf{x})] \tag{3.33}$$

του f για είσοδο $\mathbf{x}$. Η mean function που χρησιμοποιήθηκε στην παρούσα διπλωματική είναι σύνθετη $m(\mathbf{x})$ και δημιουργείται από υπάρχουσες mean functions $\mu_i(\mathbf{x})$. Συγκεκριμένα, είναι η Sum η οποία προσθέτει mean functions

$$m(\mathbf{x}) = \sum_j \mu_j(\mathbf{x}). \tag{3.34}$$

Προσθέτει την γραμμική Linear mean function, στην οποία η μέση τιμή εξαρτάται γραμμικά από το $\mathbf{x} \in X \subseteq \mathbb{R}^D$ με

$$\mu(\mathbf{x}) = a^T \cdot \mathbf{x}\ ,\ a \in \mathbb{R}^D, \tag{3.35}$$

και την σταθερή Constant mean function στην οποία η μέση τιμή ισούται με μια σταθερά,

$$\mu(\mathbf{x}) = c\ ,\ c \in \mathbb{R}. \tag{3.36}$$

Με αυτό τον τρόπο, αθροίζοντας τις (3.35) και (3.36) δημιουργείται μία συνδεδεμένη συνάρτηση (affine function) της μορφής

$$m(\mathbf{x}) = a^T \cdot \mathbf{x} + c. \tag{3.37}$$

Έτσι, με το μεν γραμμικό μέρος της συνάρτησης επιτυγχάνεται η διατήρηση της δομής του διανυσματικού χώρου των σημείων και με το δε σταθερό μέρος επιτυγχάνεται η διατήρηση της γεωμετρίας των σημείων, δηλαδή η απόσταση μεταξύ δύο σημείων.

Η covariance function $\kappa_\psi : X \times X \to \mathbb{R}$ με hyperparameter ψ, μιας Gaussian Process f είναι μία βαθμωτή συνάρτηση που ορίζεται από ολόκληρο το χώρο του X^2 και υπολογίζει την συνδιακύμανση:

$$\kappa(\mathbf{x}, \mathbf{x}') = V[f(\mathbf{x}), f(\mathbf{x}')] = E[(f(\mathbf{x}) - m(\mathbf{x})) \cdot (f(\mathbf{x}') - m(\mathbf{x}'))] \tag{3.38}$$

της f μεταξύ των εισόδων $\mathbf{x}$ και $\mathbf{x}'$. Για covariance function χρησιμοποιήθηκαν δύο μορφές συναρτήσεων, η Matern μορφή με ισοτροπικό μέτρο απόστασης (covMaterniso) και η squared exponential μορφή της covariance function (covSEiso).

Για τη Matern covariance function προκύπτει ότι:

$$X \subseteq \mathbb{R}^D, f_1(t) = 1, f_3(t) = 1 + t, f_5(t) = f_3(t) + \frac{t^2}{3}. \tag{3.39}$$

Επίσης ισχύει ότι:

$$\kappa(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \, f_d(r_d) exp(-r_d) \tag{3.40}$$

$$r_d = \sqrt{\frac{d}{\ell^2}(\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}')} \tag{3.41}$$

Η covMaterniso θεωρείται σύνθετη καθώς δέχεται σαν είσοδο μία σταθερά d που μπορεί να πάρει τις τιμές 1, 3 και 5. Η συγκεκριμένη σταθερά είναι συνυφασμένη με την ομαλότητα του Gaussian process καθώς δείχνει πόσες φορές είναι παραγωγίσιμη η συνάρτηση άρα και το πόσο ομαλή είναι.

Για τη Squared Exponential covariance function προκύπτει ότι:

$$X \subseteq \mathbb{R}^D, \;\; \kappa(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \, exp(-\frac{1}{2\ell^2}(\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}')). \tag{3.42}$$

Η covSEiso σε αντίθεση με την covMaterniso, σύμφωνα με [42, 43], είναι απείρως παραγωγίσιμη που σημαίνει πως το Gaussian process με αυτή τη συνάρτηση έχει παραγώγους όλων των τάξεων άρα η καμπύλη του είναι πολύ ομαλή.

Η likelihood function καθορίζει την πιθανότητα των παρατηρούμενων τιμών που προκύπτουν από την Gaussian process δεδομένων των hyperparameters. Στην παρούσα διπλωματική ως συνάρτηση likelihood καθορίστηκε η Gaussian συνάρτηση καθώς είναι η απλούστερη likelihood function γιατί η μεταγενέστερη διανομή της, δεν είναι απλά Gaussian αλλά μπορεί να υπολογιστεί αναλυτικά. Έτσι για την Gaussian likelihood συνάρτηση ισχύει:

$$P_\rho(y_i|f_i) = N(y_i|f_i, \sigma^2) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \; \exp\left(-\frac{(y_i - f_i)^2}{2\sigma^2}\right), \tag{3.43}$$

με *hyperparameter*, $\rho = ln\sigma$.

Ένα ακόμη χαρακτηριστικό των Gaussian processes είναι και οι Συμπερασματικές Μέθοδοι (Inference methods). Οι Inference methods καθορίζουν τον τρόπο που θα υπολογιστεί το μοντέλο, πώς να βρεθούν οι hyperparameters, ακόμη αξιολογεί το log marginal likelihood και τέλος πώς να κάνει προβλέψεις. Στην παρούσα διπλωματική χρησιμοποιήθηκε η Exact μέθοδος (infExact) για την εξαγωγή των ακριβών συμπερασμάτων.

3.6 Καταλογισμός Πλησιέστερου Γείτονα (Nearest Neighbor imputation)

Μία πολύ απλή μέθοδος που μπορεί να χρησιμοποιηθεί για την αντιμετώπιση των missing data είναι η Nearest Neighbor imputation [2, 12, 36]. Βασίζεται σε δεδομένα που υπάρχουν ήδη από παλαιότερες μετρήσεις.

Στη συγκεκριμένη μέθοδος τα δεδομένα διαχειρίζονται ως διανύσματα στήλης. Υπολογίζονται οι αποστάσεις μεταξύ του διανύσματος στο οποίο πρόκειται να γίνει καταλογισμός (imputation), για τα

σημεία που λείπουν δεδομένα, και των υπολοίπων διανυσμάτων που περιέχουν τα ήδη υπάρχοντα δεδομένα. Στη συνέχεια αντικαθίστανται τα κενά σημεία του διανύσματος που γίνεται το imputation, με τις τιμές των αντίστοιχων σημείων από το διάνυσμα που βάσει του υπολογισμού των αποστάσεων που πραγματοποιήθηκε παραπάνω, είναι ο πλησιέστερος γείτονας. Αν το αντίστοιχο σημείο του διανύσματος του πλησιέστερου γείτονα είναι επίσης κενό, τότε χρησιμοποιείται η τιμή του αμέσως επόμενου πιο κοντινού γείτονα.

Με τον αλγόριθμο που χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία, παρέχεται η δυνατότητα να αντικαθίστανται τα missing data με δεδομένα που προκύπτουν από ένα σταθμισμένο μέσο όρο των *k* πλησιέστερων γειτόνων. Τα βάρη που δίνονται στους *k* πλησιέστερους γείτονες είναι αντιστρόφως ανάλογα προς τις αποστάσεις των συγκεκριμένων διανυσμάτων από το διάνυσμα όπου επιθυμείται να γίνει το imputation.

Τέλος, ένα ακόμη χαρακτηριστικό του αλγορίθμου που χρησιμοποιήθηκε, είναι το γεγονός ότι ο υπολογισμός της απόστασης των διανυσμάτων από τον οποίο προκύπτουν οι πλησιέστεροι γείτονες μπορεί να γίνει, εκτός από τον προεπιλεγμένο τρόπο της Ευκλείδειας απόστασης, και με άλλους τρόπους σαν τους παρακάτω:

Απόσταση Οικοδομικού τετραγώνου (City Block distance)

Η απόσταση City Block ορίζεται στον $\mathbb{R}^n$ ως εξής:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|, \tag{3.44}$$

όπου x, y είναι διανύσματα του $\mathbb{R}^n$ με $x = (x_1, x_2, \ldots, x_n)$ και $y = (y_1, y_2, \ldots, y_n)$. Το όνομα της απόστασης City Block εξηγείται αν σκεφτεί κανείς δύο σημεία στο επίπεδο x-y. Η μικρότερη απόσταση των δύο σημείων είναι κατά μήκος της υποτείνουσας, η οποία είναι η Ευκλείδεια απόσταση. Η City Block απόσταση αντίθετα υπολογίζεται ως η απόσταση x συν την απόσταση y, το οποίο είναι παρόμοιο με τον τρόπο που κινείται ο κόσμος σε μια πόλη, όπου θα πρέπει να κινηθούν γύρω από τα κτίρια.

Απόσταση Chebyshev

Η απόσταση Chebyshev είναι ένα είδος μέτρησης απόστασης που ορίζεται σε ένα διανυσματικό χώρο όπου η απόσταση μεταξύ δύο διανυσμάτων είναι η μέγιστη διαφορά των συντεταγμένων των διανυσμάτων και γράφεται ως εξής:

$$d(x, y) = \max_i(|x_i - y_i|), \tag{3.45}$$

με x, y δύο διανύσματα και x_i, y_i συντεταγμένη i του κάθε διανύσματος.

Απόσταση Hamming

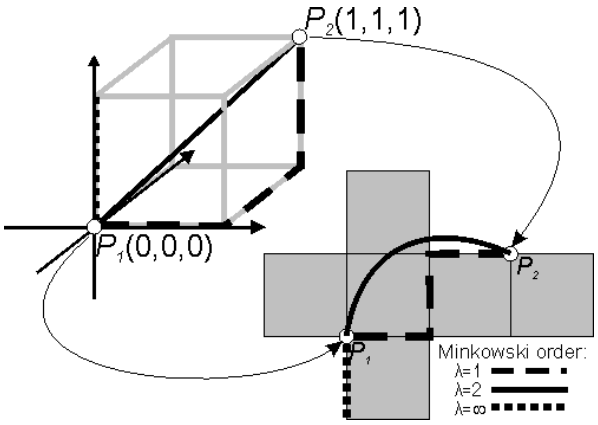
Περιγράφεται αναλυτικά στην Παράγραφο 3.10.2.

Απόσταση Minkowski

Η απόσταση Minkowski είναι μια γενικευμένη μέτρηση που περιλαμβάνει άλλα είδη μετρήσεων ως ειδικές περιπτώσεις της γενικευμένης μορφής. Η απόσταση Minkowski είναι της παρακάτω μορφής:

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^{\lambda} \right)^{\frac{1}{\lambda}}, \tag{3.46}$$

όπου x και y είναι διανύσματα μεγέθους n και λ είναι η τάξη της Minkowski μέτρησης. Παρόλο που η Συνάρτηση 3.46 ορίζεται για $\lambda > 0$, χρησιμοποιείται κυρίως για τιμές $\lambda = 1, 2$ και ∞.



Σχήμα 3.1: Αποστάσεις Minkowski για $\lambda = 1, 2, \infty$ μεταξύ δύο σημείων.[1]

Παρατηρώντας την Συνάρτηση 3.46 αλλά και με την βοήθεια του Σχήματος 3.1 προκύπτει ότι για $\lambda = 1$ η Minkowski απόσταση είναι η απόσταση City Block, για $\lambda = 2$ είναι η Ευκλείδεια απόσταση και τέλος για $\lambda = \infty$ είναι η απόσταση Chebyshev η οποία προκύπτει ως εξής:

$$d(x,y) = \lim_{\lambda \to \infty} \left(\sum_{i=1}^{n} |x_i - y_i|^{\lambda} \right)^{\frac{1}{\lambda}} = \max_i |x_i - y_i|. \tag{3.47}$$

3.7 Προσαρμογή με Νευρωνικό Δίκτυο (Neural Network fitting)

Με την χρήση του Νευρωνικού Δικτύου (Neural Network), όπως και στις προηγούμενες μεθόδους του regression, σκοπός είναι δίνοντας ένα σύνολο από ζεύγη δειγμάτων, τα σύνολα εκπαίδευσης (training sets), να βρεθεί μία συνάρτηση η οποία να ταιριάζει με αυτά.

Πιο συγκεκριμένα [15, 16], στον αλγόριθμο που χρησιμοποιείται για το Neural Network, αρχικά δημιουργείται ένα δίκτυο το οποίο είναι ένα feedforward δικτυο, δηλαδή ο πιο απλός τύπος νευρωνικού δικτύου στο οποίο δεν υπάρχουν κύκλοι ή βρόχοι αλλά οι πληροφορίες κινούνται προς μία μόνο κατεύθυνση με σκοπό την τροποποίηση ή τον έλεγχο της διαδικασίας χρησιμοποιώντας αναμενόμενα

[1] Η εικόνα στο Σχήμα 3.1 ελήφθη από τον ιστότοπο: `http://en.wikipedia.org/wiki/http://www.code10.info/index.php?option=com_content&view=article&id=61:article_minkowski-distance&catid=38:cat_coding_algorithms_data-similarity&Itemid=57`

αποτελέσματα ή τα αποτελέσματα της. Το δίκτυο αποτελείται από νευρώνες, από ένα κρυφό στρώμα (hidden layer) που χρησιμοποιεί σιγμοειδή συνάρτησή μεταφοράς η οποία είναι της μορφής:

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{3.48}$$

και ένα στρώμα εξόδου (output layer) το οποίο έχει γραμμική συνάρτηση μεταφοράς. Το πλήθος των νευρώνων και των στρωμάτων που χρησιμοποιούνται από το δίκτυο καθορίζει και την αποτελεσματικότητα του νευρωνικού δικτύου. Όσο πιο πολλοί νευρώνες και στρώματα χρησιμοποιούνται ναι μεν αυξάνουν το πλήθος των υπολογισμών αλλά επιτρέπουν στο δίκτυο να επιλύσει πολύπλοκα προβλήματα πιο αποτελεσματικά με τον μόνο περιορισμό ότι το δίκτυο έχει την τάση να παρουσιάζει υπερπροσαρμογή (overfit) στα δεδομένα όταν το πλήθος των νευρώνων είναι πολύ υψηλό. (Overfit συμβαίνει όταν ένα στατιστικό μοντέλο περιγράφει με ακρίβεια τα δεδομένα εκπαίδευσης αλλά μειώνεται η αποτελεσματικότητα εκτίμησης υπολοίπων δεδομένων εκτός των δεδομένων εκπαίδευσης.)

Στη συνέχεια εκπαιδεύεται το δίκτυο αφού όμως πρώτα χωριστούν τα δεδομένα εισόδου σε δεδομένα εκπαίδευσης, επικύρωσης και δοκιμής. Για την εκπαίδευση του δικτύου στην παρούσα διπλωματική εργασία χρησιμοποιείται ο αλγόριθμος Levenberg-Marquardt [39] και ο αλγόριθμος Bayesian regularization [9, 47].

Πιο συγκεκριμένα ο αλγόριθμος Levenberg-Marquardt ενημερώνει τις τιμές βάρους σύμφωνα με την Levenberg-Marquardt βελτιστοποίηση. Αναλυτικότερα, δεδομένου ενός συνόλου ζευγών ανεξάρτητων και εξαρτημένων μεταβλητών (x_i, y_i) ο αλγόριθμος Levenberg-Marquardt βελτιστοποιεί την παράμετρο β του μοντέλου $f(x, \beta)$ έτσι ώστε το άθροισμα των τετραγώνων των σφαλμάτων να ελαχιστοποιείται. Το άθροισμα των τετραγώνων των σφαλμάτων δίνεται από την σχέση:

$$E(\beta) = \sum_{i=1}^{m} [y_i - f(x_i, \beta)]^2 \tag{3.49}$$

Ο αλγόριθμος Levenberg-Marquardt αποτελεί μια επαναλαμβανόμενη διαδικασία. Έτσι σε κάθε βήμα επανάληψης το διάνυσμα παραμέτρου β αντικαθίσταται από μια καινούργια εκτίμηση, $\beta + \delta$. Για τον προσδιορισμό του δ οι συναρτήσεις $f(x_i, \beta + \delta)$ προσεγγίζονται ως εξής:

$$f(x_i, \beta + \delta) \approx \ f(x_i, \beta) + J_i \delta \tag{3.50}$$

όπου, $J_i = \frac{\partial f(x_i, \beta)}{\partial \beta}$ είναι η κλίση της f σε σχέση με το β. Επιπλέον χρησιμοποιούνται τα διανύσματα επικύρωσης για να σταματούν την εκπαίδευση νωρίς, αν η απόδοση του δικτύου αποτύχει να βελτιωθεί ή όταν παραμένει σταθερή για συγκεκριμένο αριθμό εκτιμήσεων που έχει οριστεί ίσο με 50.

Από την άλλη μεριά, ο αλγόριθμος Bayesian regularization ενημερώνει τα βάρη όμοια με τον προηγούμενο αλγόριθμο χρησιμοποιώντας την Levenberg-Marquardt βελτιστοποίηση, αλλά επιπλέον ελαχιστοποιεί ένα συνδυασμό των τετραγώνων των σφαλμάτων και βαρών και στη συνέχεια καθορίζει το σωστό συνδυασμό έτσι ώστε να παραχθεί ένα δίκτυο που γενικεύει καλά. Σε αντίθεση με τον προηγούμενο αλγόριθμο δεν διακόπτει την εκπαίδευση με τα διανύσματα επικύρωσης, αλλά συνεχίζει την εκπαίδευση μέχρι να βρεθεί ο βέλτιστος συνδυασμός των λαθών και βαρών.

Τέλος πρέπει να σημειωθεί, ότι κάθε φορά που ένα νευρωνικό δίκτυο εκπαιδεύεται μπορεί να αποδώσει διαφορετικά αποτελέσματα για τις ίδιες εισόδους, λόγω του ότι διαφέρουν τα αρχικά βάρη κάθε φορά που ξεκινά μια καινούργια εκπαίδευση. Για να είναι πιο ακριβή τα αποτελέσματα και να

ελαχιστοποιείται η διαφοροποίηση των αποτελεσμάτων κάθε φορά που πραγματοποιείται καινούργια εκπαίδευση, χρησιμοποιήθηκαν περισσότερα δεδομένα εισόδου. Δηλαδή θέλοντας να επιτευχθεί διόρθωση δεδομένων στην θερμοκρασία χώρου για ένα χρονικό διάστημα, χρησιμοποιήθηκαν δεδομένα εισόδου που είναι σχετικά με τον συγκεκριμένο χώρο για το ίδιο χρονικό διάστημα όπως οι τιμές της set point θερμοκρασίας και της πληρότητας (occupancy) του χώρου.

Παρόλο την δυνατότητα επιλογής διαφορετικού τρόπου υπολογισμού της απόστασης, δεν παρατηρήθηκε κάποια σημαντική διαφορά που να βελτιστοποιεί την επιλογή του πλησιέστερου γείτονα και ως εκ τούτου στην παρούσα διπλωματική, ο υπολογισμός της απόστασης πραγματοποιήθηκε υπολογίζοντας την Ευκλείδεια απόσταση.

3.8 Δείκτες Αποτελεσματικότητας των Μεθόδων

Για τον έλεγχο της αποτελεσματικότητας κάθε μεθόδου όσο αφορά την αναπαράσταση των δεδομένων και κατ' επέκταση της διόρθωσεις των δεδομένων που λείπουν, χρησιμοποιήθηκαν στην παρούσα διπλωματική οι παρακάτω τρόποι:

1. Συντελεστής προσδιορισμού (Coefficient of determination)
2. Μέσο τετραγωνικό σφάλμα (Mean Square Error)

3.8.1 Συντελεστής Προσδιορισμού (Coefficient of determination)

Στη στατιστική ο συντελεστής προσδιορισμού συμβολίζεται ως R^2 και δείχνει πόσο καλά τα δεδομένα ταιριάζουν σε ένα στατιστικό μοντέλο. Χρησιμοποιείται από στατιστικά μοντέλα των οποίων ο κύριος σκοπός είναι είτε η πρόβλεψη των αποτελεσμάτων ή η δοκιμή υποθέσεων βάσει άλλων σχετικών πληροφοριών. Ουσιαστικά παρέχει ένα μέτρο του πόσο καλά τα παρατηρούμενα αποτελέσματα αναπαράγονται από το μοντέλο [21, 53]. Ένα σύνολο δεδομένων έχει τιμές y_i, καθένα από τα οποία έχει μια αντίστοιχη τιμή από τα δεδομένα που παράχθηκαν από το μοντέλο f_i. Οι y_i τιμές καλούνται παρατηρούμενες τιμές και οι f_i τιμές καλούνται τιμές πρόβλεψης. Η «μεταβλητότητα» του συνόλου δεδομένων μετριέται μέσω διαφόρων αθροισμάτων τετραγώνων όπως τα παρακάτω:

- $SS_{tot} = \sum_i (y_i - \bar{y})^2$, το συνολικό άθροισμα των τετραγώνων
- $SS_{reg} = \sum_i (f_i - \bar{y})^2$, το άθροισμα τετραγώνων του regression
- $SS_{res} = \sum_i (y_i - f_i)^2$, το άθροισμα τετραγώνων της διαφοράς των τιμών.

Ο πιο γενικός ορισμός του συντελεστή προσδιορισμού είναι:

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}} \tag{3.51}$$

Η τιμή του R^2 μπορεί να κυμαίνεται μεταξύ 0 και 1, και όσο πιο μεγάλη είναι η τιμή του, τόσο πιο ακριβές είναι το μοντέλο του regression.

3.8.2 Μέσο Τετραγωνικό Σφάλμα (Mean Square Error)

Στη στατιστική το μέσο τετραγωνικό σφάλμα ενός εκτιμητή, μετρά τον μέσο όρο των τετραγώνων των σφαλμάτων, δηλαδή την διαφορά μεταξύ της εκτιμώμενης τιμής από την πραγματική. Η διαφορά της εκτιμώμενης τιμής από την πραγματική, παρουσιάζεται εξαιτίας της τυχαιότητας ή επειδή ο εκτιμητής δεν έχει συνυπολογίσει πληροφορίες που θα μπορούσαν να παράγουν μία ακριβέστερη εκτίμηση [53].

Αν λοιπόν θεωρηθεί το $\hat{Y}$ σαν ένα διάνυσμα με n εκτιμώμενες τιμές και Y ένα διάνυσμα με τις αντίστοιχες πραγματικές τιμές, τότε το mean square error (MSE) είναι :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{3.52}$$

Τέλος, αντίθετα με την τιμή του συντελεστή προσδιορισμού R^2, όσο πιο μικρή είναι η τιμή του MSE τόσο μικρότερο είναι το σφάλμα της εκτίμησης, άρα έχουμε αποτελεσματικότερο data correction.

3.9 Γενική Περιγραφή της Πρόβλεψης Πληρότητας ενός κτηρίου

Καταλυτικός παράγοντας στην ενεργειακή κατανάλωση των κτηρίων είναι η χρήση των συστημάτων του από τους ενοίκους του κτηρίου. Σημαντικό λοιπόν στην προσπάθεια μείωσης της ενεργειακής κατανάλωσης των κτηρίων είναι η δυνατότητα ελέγχου του κτηρίου με τέτοιο τρόπο ώστε η λειτουργία του να είναι συγχρονισμένη με την παρουσία ή μη ανθρώπων σε αυτό. Με αυτόν τον τρόπο επιτυγχάνεται σημαντική μείωση της ενέργειας που καταναλώνεται καθώς απενεργοποιούνται συστήματα του κτηρίου όταν απουσιάζουν οι κάτοικοί του. Περεταίρω μείωση της καταναλισκόμενης ενέργειας επιτυγχάνεται γνωρίζοντας το χρονικό σημείο μέσα στην ημέρα που κατοικείται εκ νέου ένα κτήριο, μετά από μία χρονική περίοδο που ηταν άδειο οπότε και τα όποια συστήματα κλιματισμού του κτηρίου ήταν ανενεργά. Έτσι δίνεται η δυνατότητα προετοιμασίας των συνθηκών του κτηρίου αμέσως πριν κατοικηθεί χρησιμοποιώντας τα συστήματα (για παράδειγμα κλιματισμού) σε χαμηλότερη ισχύ, από την πλήρη ισχύ που θα χρειαζόταν να εφαρμοστεί αν ενεργοποιούνταν τα συστήματα την ώρα που παρατηρούνταν κάτοικοι μέσα στο κτήριο.

Σε αυτή την προσπάθεια συμβάλει σημαντικά η χρήση του occupancy prediction το οποίο μας δίνει την δυνατότητα πρόβλεψης άρα και γνώσης για το πότε υφίσταται παρουσία ή όχι μέσα σε ένα κτήριο άρα επιτυγχάνεται ορθότερη διαχείριση των συστημάτων του κτηρίου άρα και σημαντική μείωση της καναλισκόμενης ενέργειας.

3.10 Μέθοδοι Πρόβλεψης Πληρότητας Κτηρίου

Υπάρχουν διάφοροι μέθοδοι πρόβλεψης πληρότητας (occupancy prediction) με αξιόλογη λειτουργία. Στην παρούσα εργασία μελετήθηκαν μέθοδοι που βασίζονται σε χρονοδιαγράμματα πληρότητας (occupancy schedules) του κτηρίου που έχουν παρατηρηθεί στο παρελθόν για ένα μεγάλο χρονικό διάστημα. Αυτές οι προσεγγίσεις αναφέρονται και ως αλγόριθμοι βασισμένοι σε χρονοδιαγράμματα

(Schedule-based algorithms). Οι μέθοδοι που αναπτύχθηκαν στηρίζονται σε εργασίες που έχουν αναπτυχθεί στο παρελθόν για την οικονομικότερη αλλά και πιο οικολογική χρήση των συστημάτων θέρμανσης μίας κατοικίας [48].

3.10.1 Αλγόριθμος SmartThermostat

Ο πρώτος αλγόριθμος που αναπτύχθηκε αναφέρεται ως SmartThermostat. Βασικό ”εργαλείο” του αλγορίθμου SmartThermostat αποτελεί το Hidden Markov Model. Έτσι κρίνεται αρχικά προτιμότερη μία συνοπτική περιγραφή του Hidden Markov Model ούτως ώστε να γίνει ευκολότερα κατανοητή η περιγραφή του αλγορίθμου SmartThermostat που ακολουθεί στην συνέχεια.

Περιγραφή του Hidden Markov Model (HMM)

Ένα HMM [1, 38] είναι ένα είδος στοχαστικού μοντέλου Markov στο οποίο παρατηρούνται μια σειρά εξόδων (emissions), αλλά δεν είναι γνωστή η ακολουθία των καταστάσεων (states) τις οποίες διέτρεξε το μοντέλο για την παραγωγή αυτών των εξόδων. Στο HMM επιδιώκεται η ανάκτηση της ακολουθίας των καταστάσεων από τα παρατηρούμενα δεδομένα. Πιο αναλυτικά, ένα HMM αποτελείται από ένα σύνολο καταστάσεων $S = S_k$ με $k = 1 : K$ να είναι το πλήθος των καταστάσεων και ένα σύνολο παρατηρούμενων εξόδων $L = V_\ell$ με $\ell = 1 : L$ να είναι το πλήθος των εξόδων. Επιπλέον, ένα HMM χαρακτηρίζεται από την πιθανότητα π_i που αποτελεί την πιθανότητα επιλογής μιας κατάστασης ως αρχική κατάσταση και ισχύει ότι $\pi_i \geq 0\ ,\ \forall\, i$ και $\sum_{i=1}^{K} \pi_i = 1$. Ακόμη, σε ένα HMM υπάρχει η πιθανότητα μεταβάσεων μεταξύ των καταστάσεων $a_{ij} = P(S_j|S_i)\ \forall i, j$ όπου ισχύει $a_{ij} \geq 0\ \forall\, i, j$ και $\sum_{j=1}^{K} a_{ij} = 1\ \forall\, i$. Τέλος η πιθανότητα να εμφανιστεί η έξοδος V_ℓ στην κατάσταση S_k είναι $b_k(\ell) = P(V_\ell|S_k)$ όπου ισχύει $b_k(\ell) \geq 0\ \forall\, k, \ell$ και $\sum_{\ell=1}^{L} b_k(\ell) = 1\ \forall\, k$.

Στην παρούσα εργασία για την χρήση του HMM, αντιμετωπίστηκαν σαν ακολουθία εξόδων τα χρονικά διαστήματα που χωρίστηκε η μέρα και σαν καταστάσεις χρησιμοποιήθηκαν δύο. Η πρώτη κατάσταση δηλώνει ότι το κτήριο κατοικείται (occupied) και η δεύτερη κατάστση που δηλώνει ότι το κτήριο δεν κατοικείται (unoccupied).

Περιγραφή αλγορίθμου SmartThermostat

Ο συγκεκριμένος αλγόριθμος αναφέρεται ως SmartThermostat (ST) λόγω του τίτλου της εργασίας [33] στην οποία έχει βασιστεί. Αρχικά πρέπει να αναφερθεί πως στις ακολουθίες εξόδων, δηλαδή τα χρονικά διαστήματα, και στις ακολουθίες καταστάσεων, δηλαδή 1 για occupied ή 0 για unoccupied, χρησιμοποιήθηκαν δεδομένα του κτηρίου που έχουν καταγραφεί στο παρελθόν για ένα μεγάλο χρονικό διάστημα. Οι παραπάνω ακολουθίες χρησιμοποιήθηκαν ως δεδομένα εκπαίδευσης του Hidden Markov μοντέλου.

Ως πρώτο βήμα του αλγορίθμου είναι ο υπολογισμός δύο πινάκων, του πίνακα εξόδων και του πίνακα μεταβάσεων. Στον μεν πίνακα εξόδων περιέχονται οι πιθανότητες που έχει η κάθε έξοδος που αντιστοιχεί σε μία χρονική περίοδο της ημέρας, ώστε να παρατηρηθεί. Στο δε πίνακα μεταβάσεων

περιέχονται οι πιθανότητες της κάθε μετάβασης από την μία κατάσταση στην άλλη ή την παραμονή στην ίδια κατάσταση.

Έπειτα, δεδομένου ότι έχουν υπολογιστεί οι πίνακες μεταβάσεων και εξόδων βάσει των δεδομένων εκπαίδευσης, ο αλγόριθμος υπολογίζει την πιο πιθανή ακολουθία καταστάσεων που προκύπτει δίνοντας στο εκπαιδευμένο πλέον HMM μια δεδομένη ακολουθία εξόδων. Στην προκειμένη περίπτωση η ακολουθία εξόδων βάσει της οποίας προκύπτει η ακολουθία των καταστάσεων, είναι το χρονικό διάστημα της ημέρας που θέλουμε να προβλέψουμε το occupancy.

Ο υπολογισμός της πιο πιθανής ακολουθίας καταστάσεων προκύπτει με την χρήση του αλγορίθμου Viterbi. Ο αλγόριθμος Viterbi λειτουργεί ώς εξής:

Υποθέτοντας την ύπαρξη ενός Hidden Markov Model με χώρο καταστάσεων *S*, αρχική πιθανότητα π_i να βρίσκεται στην κατάσταση i και πιθανότητα $\alpha_{i,j}$ να μεταβεί απο την κατάσταση i στην κατάσταση j. Έστω ότι παρατηρούμε εξόδους $y_1, y_2, ..., y_T$. Η πιο πιθανή ακολουθία καταστάσεων $x_1, x_2, ..., x_T$ που παράγει τις παρατηρούμενες τιμές δίνεται απο τις εξής σχέσεις:

$$V_{1,k} = P(y_1|k) \times \pi_k \tag{3.53}$$

$$V_{t,k} = P(y_t|k) \times \max_{x \in S}(\alpha_{x,k} \times V_{t-1,x}) \tag{3.54}$$

Εδώ, $V_{t,k}$ είναι η πιθανότητα της πιο πιθανής ακολουθίας καταστάσεων που είναι υπεύθυνη για τις πρώτες *t* παρατηρούμενες τιμές που προκύπτουν όταν το μοντέλο βρίσκεται στην *k* κατάσταση. Η διαδρομή Viterbi μπορεί να ανακτηθεί αποθηκεύοντας δείκτες που θυμούνται ποιά κατάσταση *x* χρησιμοποιήθηκε στην συνάρτηση 3.54. Έστω ότι $Ptr(k, t)$ είναι η συνάρτηση που επιστρέφει την τιμή του *x* που χρησιμοποιήθηκε για τον υπολογισμό του $V_{t,k}$ αν $t > 1$, ή *k* αν *t=1*. Τότε:

$$x_T = \arg\max_{x \in S}(V_{T,x}) \tag{3.55}$$

$$x_{t-1} = Ptr(x_t, t) \tag{3.56}$$

Εδώ χρησιμοποιείται ο βασικός ορισμός του arg max, δηλαδή ότι, το $\arg\max_x f(x)$ είναι το σύνολο των τιμών του x για τις οποίες η $f(x)$ φθάνει τη μέγιστη τιμή της.

3.10.2 Περιγραφή αλγορίθμου PreHeat

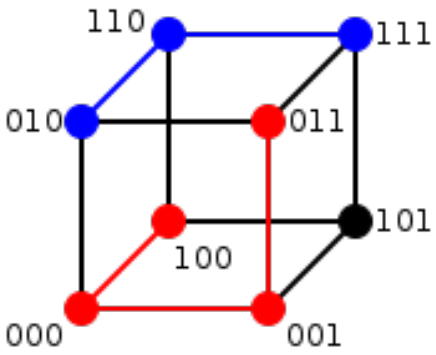
Σε αυτή την παράγραφο, ως συνέχεια της περιγραφής των μεθόδων πρόβλεψης πληρότητας ενός κτηρίου, παρουσιάζεται ο αλγόριθμος PreHeat (PH) του οποίου το όνομα προκύπτει από τον τίτλο της εργασίας [48] που έχει βασιστεί. Ο συγκεκριμένος αλγόριθμος αναπαριστά την παρουσία κατοίκων σε ένα κτήριο σε συγκεκριμένα χρονικά διαστήματα μέσα στην ημέρα με την χρήση ενός δυαδικού διανύσματος. Δηλαδή, όταν στο κτήριο υπάρχει έστω και ένας κάτοικος τότε το διάνυσμα παίρνει την τιμή 1. Αντίστοιχα όταν δεν υπάρχει κανείς εντός του κτηρίου το διάνυσμα παίρνει την τιμή 0.

Πιο αναλυτικά, χρησιμοποιεί ένα διάνυσμα όπου καταχωρείται η πληρότητα (occupancy) της τρέχουσας ημέρας. Γνωρίζοντας μέχρι ένα χρονικό σημείο την κατάστηση πληρότητας του κτηρίου, ο αλγόριθμος προβλέπει την κατάσταση πληρότητας για το υπόλοιπο της ημέρας. Για να επιτευχθεί αυτό, ο αλγόριθμος αρχικά υπολογίζει την απόσταση Hamming (περιγράφεται στην συνέχεια) μεταξύ των ήδη γνωστών δεδομένων της τρέχουσας ημέρας και δεδομένων προηγούμενων ημερών που αντιστοιχούν στην ίδια χρονική περίοδο με τα γνωστά δεδομένα της τρέχουσας ημέρας. Έπειτα υπολογίζει

την πιθανότητα του προβλεπόμενου occupancy για κάθε χρονική στιγμή, ως την μέση τιμή των αντίστοιχων τιμών occupancy από τα k πιο κοντινά διανύσματα των προηγούμενων ημερών. Σύμφωνα με την εργασία [48] το k ισούται με την τιμή 5, καθώς έχει αποδειχθεί μέσα από πειράματα ως μία πολύ καλή επιλογή για υψηλή ακρίβεια πρόβλεψης. Τέλος βάσει ενός ορίου, στην πιθανότητα να παρατηρηθεί occupancy, που ορίζεται από τον χρήστη ανάλογα με τις απαιτήσεις του, καθορίζεται η τελική απόφαση για το occupancy. Δηλαδή όταν η πιθανότητα ξεπερνά το όριο που έχει τεθεί τότε προβλέπεται ότι κάποιος υπάρχει μέσα στο κτήριο και άρα καταχωρείται η τιμή 1 στο αντίστοιχο διάνυσμα, στην αντίθετη περίπτωση καταχωρείται η τιμή 0 στο διάνυσμα.

Απόσταση Hamming

Ως απόσταση Hamming μεταξύ δύο συμβολοσειρών ίσου μήκους ορίζεται ο αριθμός θέσεων στις οποίες τα αντίστοιχα σύμβολα είναι διαφορετικά. Η απόσταση Hamming, μετρά τον ελάχιστο αριθμό αντικαταστάσεων που χρειάζονται ώστε να μετατραπεί η μία συμβολοσειρά στην άλλη ή αλλιώς, τον αριθμό των λαθών που μετέτρεψαν την μια συμβολοσειρά στην άλλη. Για παράδειγμα η απόσταση Hamming μεταξύ του *1011101* και *1001001* είναι 2.



Σχήμα 3.2: Κύβος δυαδικών αριθμών με 3 ψηφία για την εύρεση της απόστασης Hamming.[2]

Στο Σχήμα 3.2 παρουσιάζονται δύο επιπλέον παραδείγματα της απόστασης Hamming. Στο παράδειγμα με την κόκκινη γραμμή παρουσιάζεται η απόσταση Hamming μεταξύ του *100* και *011* που έχουν απόσταση Hamming 3, και στο παράδειγμα με την μπλε γραμμή παρουσιάζεται η απόσταση Hamming μεταξύ του *010* και *111* που έχουν απόσταση Hamming 2.

[2]Η εικόνα στο Σχήμα 3.2 ελήφθη από τον ιστότοπο: `http://en.wikipedia.org/wiki/Hamming_distance`

Κεφάλαιο 4

Πειράματα και Παράθεση Αποτελεσμάτων των Μεθόδων Διόρθωσης Δεδομένων

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν εκτενώς σε θεωρητικό πλαίσιο οι μέθοδοι που μελετήθηκαν στην παρούσα διπλωματική εργασία και αφορούν τόσο την διόρθωση των δεδομένων που λείπουν όσο και την πρόβλεψη πληρότητας ενός κτηρίου.

Στην συνέχεια σε αυτό το κεφάλαιο παρουσιάζονται τα πειράματα και τα αποτελέσματα της εφαρμογής των μεθόδων διόρθωσης δεδομένων πάνω σε πραγματικά δεδομένα κτηρίων σε μία προσπάθεια κατανόησης της συμπεριφοράς και της αποτελεσματικότητας των μεθόδων αυτών σε όσο πιο ρεαλιστικές συνθήκες μπορούν να προσφέρουν τα πραγματικά δεδομένα κάνοντας τις λιγότερες δυνατές υποθέσεις.

Τα πειράματα πραγματοποιήθηκαν εκτελώντας τις μεθόδους με την βοήθεια του μαθηματικού εργαλείου Matlab, το οποίο προσφέρει δυνατότητα επεξεργασίας των μεθόδων χρησιμοποιώντας μεγάλο πλήθος δεδομένων και διευκολύνει την εξαγωγή συμπερασμάτων μέσω των γραφικών παραστάσεων που μπορούν να σχεδιαστούν.

4.1 Δεδομένα που χρησιμοποιήθηκαν

4.1.1 Κτήριο μετρήσεων

Για την εξαγωγή συμπερασμάτων σχετικά με την αποτελεσματικότητα των μεθόδων διόρθωσης δεδομένων και των μεθόδων πρόβλεψης πληρότητας που μελετήθηκαν στην παρούσα εργασία ήταν διαθέσιμα δεδομένα μετρήσεων θερμοκρασίας χώρου, set point θερμοκρασιών και δεδομένα πληρότητας (occupancy) του χώρου, και πιο συγκεκριμένα ενός γραφείου του κτηρίου των συνεργατών του Ευρωπαϊκού προγράμματος BaaS [11] στην Ισπανία.

Πιο συγκεκριμένα το κτήριο αποτελεί το τεχνολογικό κέντρο CARTIF [26] , το οποίο ουσιαστικά είναι κτήριο γραφείων που κατασκευάστηκε το 1995 και βρίσκεται λίγο έξω από την Valladolid της Ισπανίας σε υψόμετρο 720 μέτρων. Το κτήριο αποτελείται από τρεις ορόφους (υπόγειο, ισόγειο και πρώτος όροφος) που περιέχουν κυρίως γραφεία αλλά και άλλους χώρους. Στο υπόγειο όπου βρίσκο-

νται χώροι όπως το λεβητοστάσιο και αποθήκες, δεν υπάρχει θέρμανση ή κλιματισμός. Οι δύο επόμενοι όροφοι, δηλαδή το ισόγειο και ο πρώτος όροφος, αποτελούνται από γραφεία και εργαστήρια.

Σχήμα 4.1: Εξωτερική όψη του κτηρίου CARTIF[1]

Για τον κλιματισμό των διαφόρων χώρων στο ισόγειο και στον πρώτο όροφο χρησιμοποιούνται ολοκληρωμένα συστήματα όπως ενεργές θερμικές πλάκες (thermal active slabs) που καλύπτουν το σύνολο της επιφάνειας των δύο αυτών ορόφων, αντλίες θερμότητας που χρησιμοποιούν νερό (water source heat pumps) και είναι τοποθετημένες σε ορισμένα εργαστήρια, αερόθερμα και fan coils που παρέχουν και θέρμανση και ψύξη και τέλος θερμαντικά σώματα (convective radiators) που χρησιμοποιούνται μόνο σε χώρους που έχουν υψηλότερες απαιτήσεις θέρμανσης. Το συγκεκριμένο κτήριο έχει μέση κατανάλωση ενέργειας $128,06\ kWh/m^2$ το χρόνο και μέση θερμική κατανάλωση $55,75\ kWh/m^2$ το χρόνο. Πιο αναλυτικά καταναλώνει συνολικά $339.411,0\ kWh$ ηλεκτρικής ενέργειας και $144.554,25\ kWh$ ενέργεια από την καύση φυσικού αερίου.

Για την αποτελεσματικότερη παρακολούθηση αλλά και έλεγχο του κτηρίου της CARTIF, έχει επιλεχθεί η τεχνολογία του δικτύου LonWorks® , το οποίο αποτελείται από κόμβους που επικοινωνούν μεταξύ τους μέσω του πρωτοκόλλου LonTalk®. Το συγκεκριμένο δίκτυο περιλαμβάνει τις ηλιακές εγκαταστάσεις (θερμικούς ηλιακούς συλλέκτες, ενεργές θερμικές πλάκες, λέβητα φυσικού αερίου, και αντλίες θερμότητας), το φωτοβολταϊκό πεδίο, το φωτισμό και το σύστημα πρόσβασης του κτηρίου. Υποσυστήματα του δικτύου είναι διάφοροι ελεγκτές, πομποδέκτες, συσκευές ελέγχου του δικτύου, διεπαφές που επιτρέπουν την σύνδεση συσκευών όπως ηλεκτρονικοί υπολογιστές και PLCs. Ακόμη το δίκτυο περιλαμβάνει εργαλεία για την εγκατάσταση, τη διαμόρφωση και τη διάγνωση των κόμβων του και τέλος εργαλεία για την ανάπτυξη εφαρμογών.

Όλα αυτά έχουν σαν αποτέλεσμα κάθε χώρος και όλα τα στοιχεία του να παρακολουθούνται. Παρέχεται η δυνατότητα ελέγχου των μεταβλητών και της κατάστασης του κτηρίου γενικότερα, μέσω ειδικών εφαρμογών που επιτρέπουν την πρόσβαση στο σύστημα μέσω Ethernet, internet και e-mail.

[1]Η εικόνα βρίσκεται στο [26].

Με αυτόν τον τρόπο οι στρατηγικές διαχείρισης του κτηρίου μπορούν να αλλάξουν με ευκολία ώστε να προσαρμοστούν στις εκάστοτε συνθήκες, και επιπλέον τυχόν σφάλματα στις εγκαταστάσεις μπορούν γρήγορα να ανιχνευθούν και να λυθούν.

Τέλος αξίζει να σημειωθεί ότι προκειμένου να αποθηκεύονται όλα τα δεδομένα από το σύστημα διαχείρισης του κτηρίου συμπεριλαμβανομένων και των παραμέτρων που περιλαμβάνονται στο δίκτυο LonWorks® χρησιμοποιείται η βάση δεδομένων PostgreSQL® η οποία είναι ανοιχτού κώδικα (open source) και είναι διαθέσιμη για όλα τα βασικά λειτουργικά συστήματα.

4.1.2 Περιγραφή των Δεδομένων

Οι πηγές δεδομένων σύμφωνα με το σύστημα BaaS [27] συνοψίζονται παρακάτω ως εξής:

- Το μοντέλο πληροφοριών κτηρίου (Building Information Model(BIM)) που περιέχεται στο BIM Server και περιλαμβάνει όλες τις στατικές πληροφορίες σχετικά με το κτήριο όπως τη γεωμετρία του, τα υλικά κατασκευής του, τα συστήματα του κ.ο.κ.

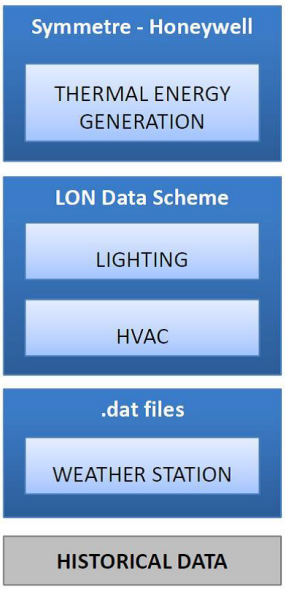
- Την αποθήκη δεδομένων (Data Warehouse), που περιέχει όλες τις δυναμικές πληροφορίες του κτηρίου συμπεριλαμβανομένων των ιστορικών δεδομένων και των απευθείας δεδομένων.

- Το σύστημα διαχείρισης κτιρίου (Building Management System) το οποίο επιτρέπει την πρόσβαση στα απευθείας δεδομένα και τον έλεγχο των ενεργειακών συστημάτων.

- Εξωτερικές πηγές δεδομένων που συνδέονται με το κτήριο όπως μετεωρολογικοί σταθμοί, συστήματα ελέγχου πρόσβασης και άλλα.

- Εξωτερικές υπηρεσίες, όπως υπηρεσίες πρόγνωση καιρού.

Τα ιστορικά δεδομένα που ενδιαφέρουν την παρούσα διπλωματική εργασία περιέχονται στην αποθήκη δεδομένων (Data Warehouse) που όσο αφορά το κτήριο της CARTIF απαρτίζεται από τρεις πηγές δεδομένων. Την βάση δεδομένων του δικτύου LonWorks®, τα αρχεία csv για τα δεδομένα του εξωτερικού μετεωρολογικού σταθμού και τη βάση δεδομένων που χρησιμοποιεί το SymmetrE. Στο Σχήμα 4.2 παρουσιάζεται σχηματικά η κατανομή των ιστορικών δεδομένων που είναι διαθέσιμα στο κτίριο της CARTIF.



Σχήμα 4.2: Οι υπάρχουσες αποθήκες δεδομένων στο CARTIF[2]

Πιο αναλυτικά [27], στο δίκτυο αισθητήρων LonWorks® [23] οι συσκευές συνδέονται σε μία πύλη που ονομάζεται iLon [22]. Αυτή η συσκευή αποθηκεύει προσωρινά τα δεδομένα σε αρχεία csv σε μία εσωτερική περιορισμένη μνήμη. Έπειτα αυτά τα δεδομένα αποθηκεύονται σε μια μηνιαία βάση δεδομένων η οποία περιέχει τέσσερις πίνακες, ένας για κάθε εβδομάδα του μήνα. Αυτό συμβαίνει λόγω της μεγάλης ποσότητας των δεδομένων που παρακολουθούνται στη βάση δεδομένων. Πάνω από ένας μήνας δεδομένων δεν είναι διαχειρίσιμος σε μια βάση δεδομένων, επειδή η απόδοση μειώνεται εκθετικά. Στη συνέχεια δημιουργούνται αντίγραφα ασφαλείας και γίνεται καθαρισμός της βάσης δεδομένων για το νέο μήνα για λόγους αποθήκευσης των νέων δεδομένων. Έτσι, μια σειρά από μηνιαία αρχεία αντιγράφων ασφαλείας είναι διαθέσιμα ως ιστορικά δεδομένα. Τα αρχεία αυτά αποθηκεύονται στη βάση δεδομένων PostgreSQL®.

Τα δεδομένα του εξωτερικού μετεωρολογικού σταθμού δεν καταχωρούνται σε βάσεις δεδομένων αλλά σε αρχεία δεδομένων. Τα αρχεία αυτά αντιγράφονται καθημερινά ώστε να δημιουργηθεί ένα ιστορικό των τιμών του εξωτερικού μετεωρολογικού σταθμού. Αυτά τα αρχεία καταγραφής οργανώνονται σε στήλες που περιέχουν τις πληροφορίες των αρχείων.

Τέλος υπάρχει το SymmetrE [29] ένας σταθμός εργασίας για τα συστήματα θερμικής παραγωγής ενέργειας. Αυτή η συσκευή είναι ένα ρυθμιζόμενο σύστημα διαχείρισης το οποίο είναι σε θέση όχι μόνο να διαβάσει τις μεταβλητές και τις τιμές από το δίκτυο, αλλά επίσης και να ελέγχει τα συστήματα και τις εγκαταστάσεις. Το SymmetrE αποθηκεύει τις πληροφορίες σε μια τοπική βάση δεδομένων διαθέσιμη σε πραγματικό χρόνο η οποία είναι ο Microsoft SQL Server [34]. Αυτή η συσκευή είναι σε θέση να παράγει, αρχεία Excel και ένα ιστορικό μητρώο που βασίζεται στο Microsoft Access με τις ιστορικές καταγραφές.

Από τις συγκεκριμένες πηγές δεδομένων είναι πλέον διαθέσιμα τα δεδομένα για περαιτέρω επεξεργασία και ανάλυση που στην παρούσα διπλωματική εργασία πραγματοποιήθηκε χρησιμοποιώντας

[2]Η εικόνα βρίσκεται στο [27].

το Matlab. Η καταγραφή των δεδομένων αυτών γίνεται σε χρονικά διαστήματα που εμφανίζονται μεταβολές στις τιμές των αισθητήρων, δηλαδή μόλις αλλάξει η τιμή της θερμοκρασίας ή η τιμή του occupancy του κτηρίου τότε το σύστημα καταγράφει αυτή την τιμή.

Για να επιτευχθεί καλύτερη και λεπτομερέστερη διαχείριση των δεδομένων σε όλο το φάσμα του 24ώρου, πραγματοποιήθηκε interpolation στα δεδομένα με σκοπό να υπάρχει σταθερό interval μεταξύ των τιμών, το οποίο ορίστηκε στα πέντε λεπτά. Έπειτα διαχωρίστηκαν τα δεδομένα σε καθημερινές μέρες και μέρες Σαββατοκύριακου, λόγω του ότι στο γραφείο τα Σαββατοκύριακα δεν παρατηρείται παρουσία ατόμων οπότε και οι τιμές των δεδομένων διαφοροποιούνται σε σχέση με αυτές των καθημερινών ημερών.

Τα πειράματα που παρουσιάζονται παρακάτω αφορούν διαφορετικές μέρες ή σύνολα ημερών από το σύνολο των δεδομένων που είναι διαθέσιμα. Στα συγκεκριμένα δεδομένα που ήταν διαθέσιμα δεν υπήρχαν missing data και ως εκ τούτου για να πραγματοποιηθούν τα πειράματα αφαιρέθηκαν κάποια δεδομένα με σκοπό να υπάρξουν missing data και άρα να μπορεί να μελετηθεί η συμπεριφορά των μεθόδων διόρθωσης δεδομένων.

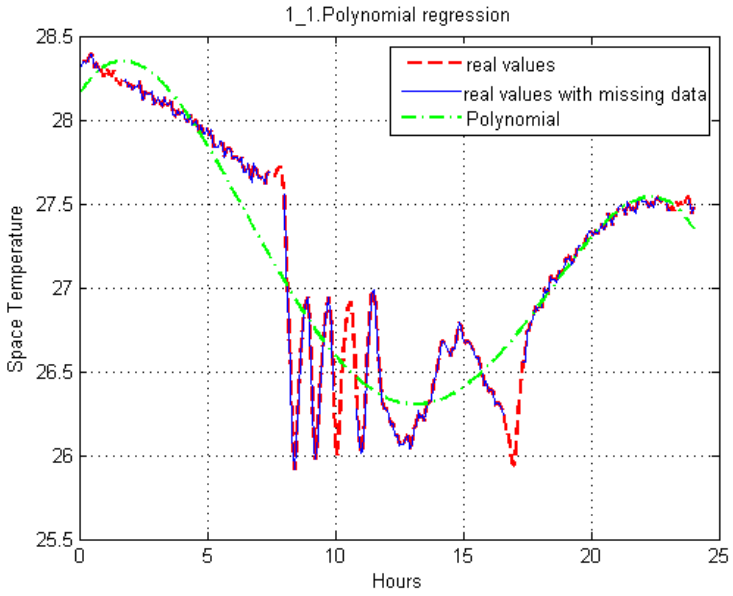
4.2 Πείραμα 1

Στο συγκεκριμένο πείραμα επιλέχθηκαν τα δεδομένα θερμοκρασίας μίας ημέρας, από το σύνολο των διαθέσιμων δεδομένων, με κύριο κριτήριο τις έντονες αυξομειώσεις της θερμοκρασίας κατά την διάρκεια της συγκεκριμένης ημέρας. Στις γραφικές παραστάσεις που ακολουθούν αναπαριστάται η εφαρμογή των μεθόδων που παρουσιάστηκαν στο προηγούμενο κεφάλαιο. Σε αυτό το δείγμα θερμοκρασιών του γραφείου μίας ολόκληρης μέρας, λείπουν δεδομένα περίπου 3 ωρών και 30 λεπτών.

Στα σχήματα που ακολουθούν, όπως φαίνεται και στα υπομνήματα, με την κόκκινη διακεκομμένη γραμμή αναπαριστώνται τα πραγματικά δεδομένα ενώ με την μπλε γραμμή αναπαριστώνται τα ίδια δεδομένα αλλά με ελλείποντες τιμές. Επιπλέον με την πράσινη γραμμή παρουσιάζονται τα αποτελέσματα της κάθε μεθόδου. Οι τιμές που λείπουν έχουν δημιουργηθεί στο Matlab, εκχωρώντας στις θέσεις που υπάρχουν τα κενά την τιμή *NaN*.

Polynomial Regression

Στο Σχήμα 4.3 παρουσιάζεται η μέθοδος Polynomial Regression που εφαρμόστηκε στα δεδομένα του Πειράματος 1. Η επιλογή του πολυωνύμου που χρησιμοποιείται από την μέθοδο του Polynomial regression γίνεται βάσει δοκιμών που πραγματοποιήθηκαν στη Matlab με σκοπό την όσο δυνατόν καλύτερη προσέγγιση του regression στα πραγματικά δεδομένα. Γενικά ισχύει ότι αυξάνοντας το βαθμό του πολυωνύμου βελτιώνεται η αποτελεσματικότητα της μεθόδου διότι αυξάνονται και οι όροι των παραμέτρων α που αναφέρονται στην Παράγραφο 3.2. Βέβαια δεν είναι πάντα αποτελεσματικό να χρησιμοποιείται ο κατά το δυνατόν μεγαλύτερος βαθμός πολυωνύμου καθώς όπως εξηγείται και στο [20] ναι μεν επιτυγχάνεται βέλτιστη προσέγγιση των δεδομένων εκπαίδευσης του μοντέλου αλλά από ένα βαθμό πολυωνύμου (ανάλογα με τα δεδομένα) και έπειτα μειώνεται η αποτελεσματικότητα προσέγγισης δεδομένων εκτός των δεδομένων εκπαίδευσης με αποτέλεσμα να χάνεται η γενικότερη εικόνα των δεδομένων και ουσιαστικά να αναπαριστώνται μόνο τα δεδομένα εκπαίδευσης. Στο συγκεκριμένο πείραμα το πολυώνυμο που χρησιμοποιείται είναι $5^{ου}$ βαθμού. Με την επιλογή αυτού του βαθμού για το πολυώνυμο γίνεται μία αποτελεσματικότερη εκτίμηση του συνόλου των δεδομένων, δηλαδή εκτός



Σχήμα 4.3: Στην παραπάνω γραφική παράσταση το πολυώνυμο που χρησιμοποιείται είναι 5ου βαθμού.

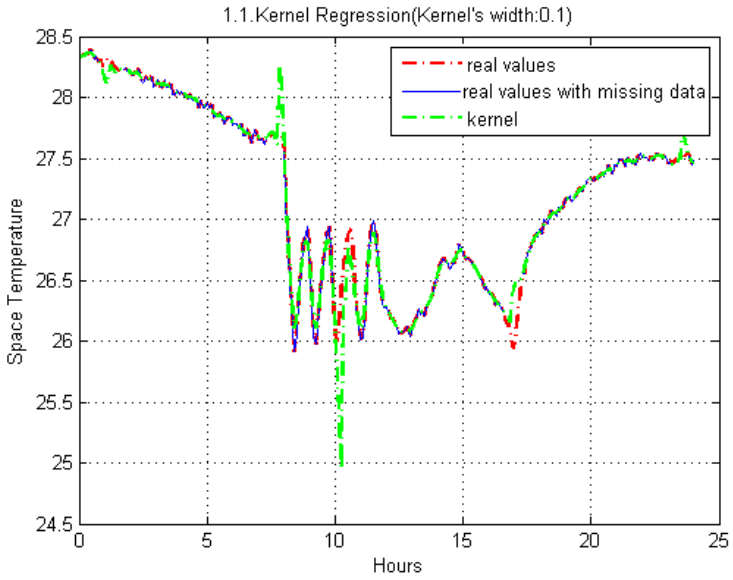
από τα ήδη γνωστά δεδομένα που χρησιμοποιούνται ως δεδομένα εκπαίδευσης προσεγγίζονται αποτελεσματικότερα και τα ελλείποντα δεδομένα. Ωστόσο στο συγκεκριμένο πείραμα παρά την προσπάθεια πιο γενικευμένης προσέγγισης των δεδομένων η μέθοδος του Polynomial Regression αποτυγχάνει καθώς οι έντονες αυξομειώσεις των δεδομένων καθιστούν πολύ δύσκολη την λεπτομερέστερη πρόγνωση των τιμών τους.

Local(Kernel) Regression

Με τα Σχήματα 4.4, 4.5 και 4.6, που παρουσιάζονται στη συνέχεια φαίνεται η αποτελεσματικότητα της μεθόδου Local Regression σχετικά με τα δεδομένα του Πειράματος 1.

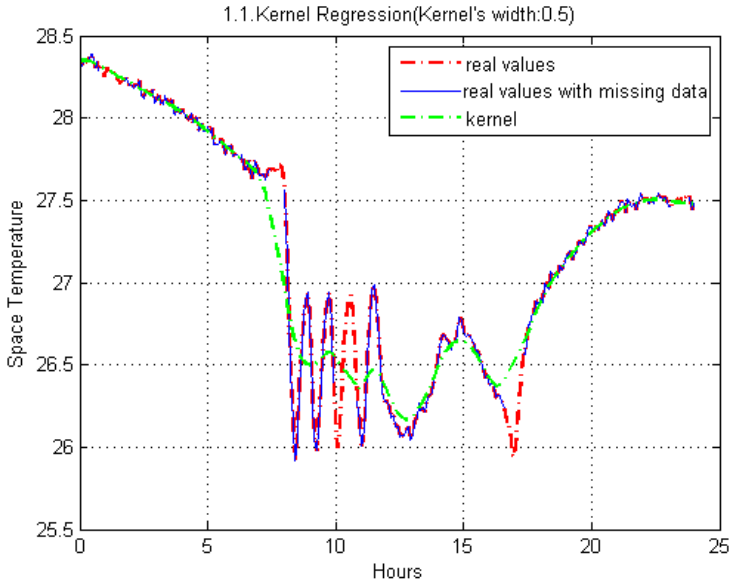
Η συγκεκριμένη μέθοδος παρέχει την δυνατότητα επιλογής της παραμέτρου εξομάλυνσης, h, που περιγράφεται στην Παράγραφο 3.3. Εφόσον η παράμετρος εξομάλυνσης είναι μικρή η εκτίμηση του Local Regression παρακολουθεί πιο στενά τα δεδομένα, δηλαδή προβλέπει με μεγαλύτερη ακρίβεια τις τιμές των δεδομένων. Αυτό συμβαίνει διότι η μέθοδος χρησιμοποιεί ως δεδομένα εκπαίδευσης όλο και πιο κοντινά δεδομένα στο σημείο πρόβλεψης. Ακριβώς αντίθετα, όσο μεγαλώνει η παράμετρος εξομάλυνσης η εκτίμηση γίνεται πιο γενική. Το πόσο μικρή ή μεγάλη μπορεί να θεωρηθεί η τιμή της παραμέτρου εξομάλυνσης έχει να κάνει αποκλειστικά με τις ιδιότητες των δεδομένων που χρησιμοποιούνται όπως περιγράφεται στην Παράγραφο 3.3 και πιο συγκεκριμένα στην Συνάρτηση 3.11.

Στο Σχήμα 4.4 η τιμή της παραμέτρου εξομάλυνσης ισούται με $h = 0.1$ και σταδιακά στα επόμενα δύο σχήματα η παράμετρος εξομάλυνσης αυξάνεται ως την τιμή $h = 1$. Αρχικά παρατηρείται ότι όσο πιο μικρή είναι η τιμή της παραμέτρου τόσο πιο λεπτομερής είναι η εκτίμηση της μεθόδου ως προς τα δεδομένα, προς επιβεβαίωση των όσων προαναφέρθηκαν. Αυξάνοντας την τιμή της παραμέτρου στο 0.5, γίνεται αντιληπτό ότι η μέθοδος συνεχίζει να εκτιμά με ακρίβεια τα σημεία όπου η πυκνότητα των δεδομένων είναι μικρή ενώ αρχίζει να κάνει μια πιο γενική εκτίμηση εκεί όπου τα δεδομένα είναι πιο



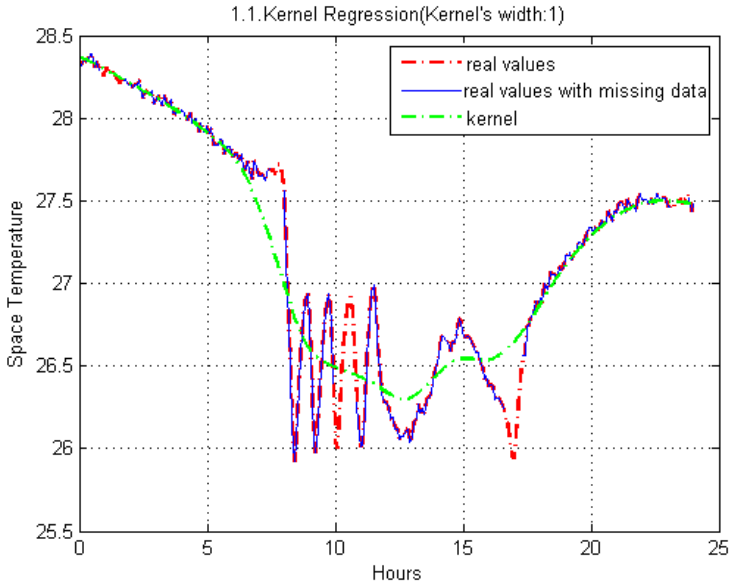
Σχήμα 4.4: Επιλογή της παραμέτρου εξομάλυνσης h=0.1

πυκνά όπως φαίνεται στο Σχήμα 4.5 λίγο πριν τις 10:00. Αυτό οφείλεται στο γεγονός ότι αυξάνοντας την τιμή της παραμέτρου εξομάλυνσης χρησιμοποιούνται δεδομένα από μία ευρύτερη περιοχή και λόγω του ότι στην συγκεκριμένη περιοχή τα δεδομένα είναι περισσότερα και παρουσιάζουν έντονη μεταβολή στην τιμή τους έχει σαν αποτέλεσμα η εκτίμηση της μεθόδου να μην είναι ακριβής.

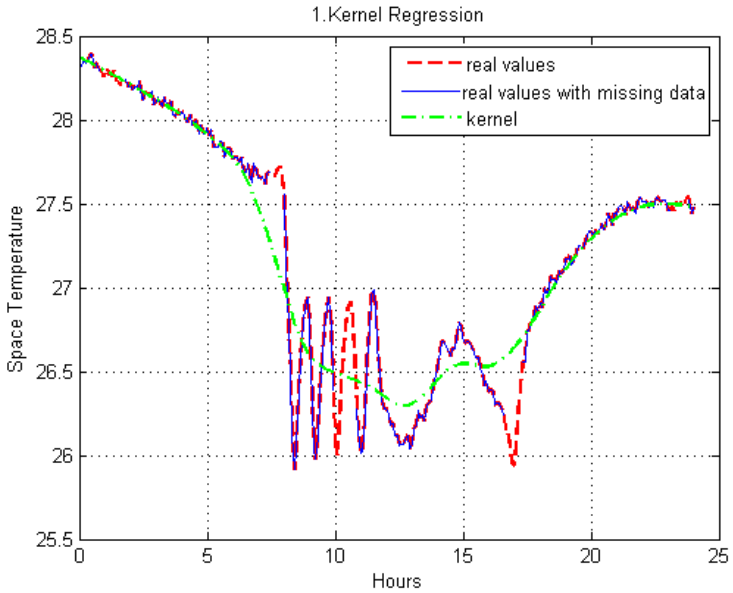


Σχήμα 4.5: Επιλογή της παραμέτρου εξομάλυνσης h=0.5

Στο Σχήμα 4.7 παρουσιάζεται το αποτέλεσμα του Local Regression χρησιμοποιώντας την επιλογή



Σχήμα 4.6: Επιλογή της παραμέτρου εξομάλυνσης h=1

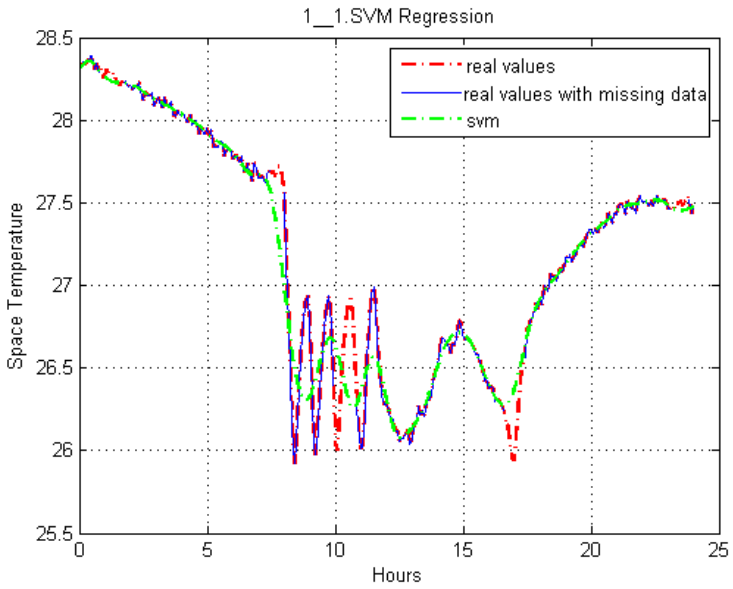


Σχήμα 4.7: Χρήση του τρόπου που περιγράφεται στην παράγραφο 3.3 για τον υπολογισμό της παραμέτρου εξομάλυνσης. Σε αυτή την περίπτωση προκύπτει h=0.9747

του αλγορίθμου να υπολογίζει τη βέλτιστη παράμερο εξομάλυνσης σύμφωνα με τον τρόπο που περιγράφτηκε στην Παράγραφο 3.3. Η μεγάλη τιμή του h που επιλέγεται από τον αλγόριθμο σε αυτή της περίπτωση βάσει του βιβλίου [21], οφείλεται στο γεγονός ότι η προσέγγιση που ακολουθείται είναι συντηρητική και βασίζεται στην κανονική κατανομή η οποία λόγω του ότι είναι μία από τις ομαλότερες κατανομές έχει σαν αποτέλεσμα να προκύπτει μεγάλη τιμή για την βέλτιστη παράμετρο εξομάλυνσης. Επιπλέον όταν χρησιμοποιούνται μη κανονικά δεδομένα το αποτέλεσμα της βέλτιστης παραμέτρου εξομάλυνσης οδηγεί σε υπερ-ομαλές εκτιμήσεις.

Έτσι λοιπόν λόγω του ότι τα δεδομένα στο Πείραμα 1 αποκλίνουν αρκετά από την κανονική κατανομή, οδηγούν την εκτίμηση της μεθόδου στο Σχήμα 4.7 να είναι γενικευμένη και να μην προσεγγίζει με λεπτομέρεια τα δεδομένα καθώς ως βέλτιστη τιμή της παραμέτρου εξομάλυνσης προκύπτει μεγάλη τιμή για αυτά τα δεδομένα.

Support Vector Regression

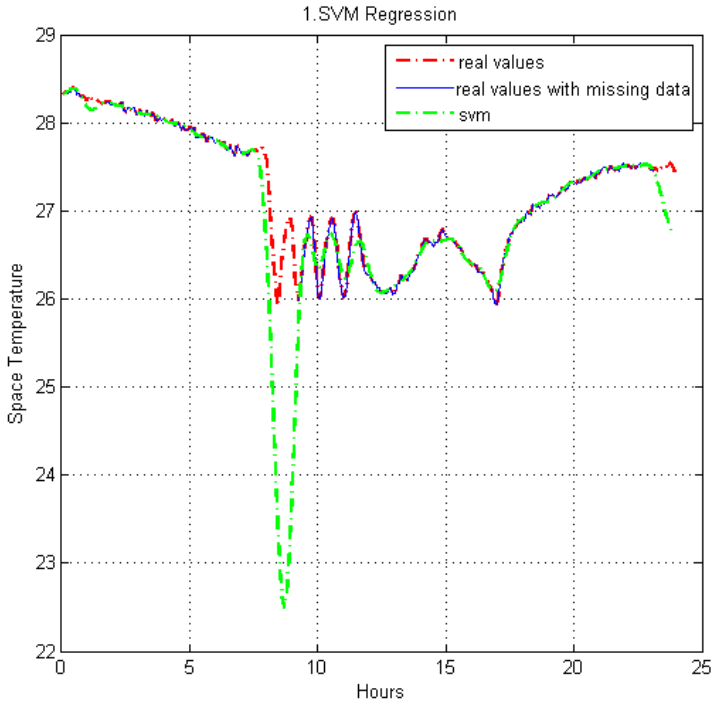


Σχήμα 4.8: Τα δεδομένα, λείπουν σε μικρά κομμάτια κατά την διάρκεια του 24ώρου.

Στα Σχήματα 4.8, 4.9 εμφανίζεται το αποτέλεσμα της εφαρμογής της μεθόδου Support Vector Regression στα δεδομένα του Πειράματος 1. Στον αλγόριθμο που χρησιμοποιήθηκε σε αυτή τη μέθοδο η ακρίβεια του αποτελέσματος εξαρτάται άμεσα από την επιλογή των κατάλληλων παραμέτρων C, ν και γ που περιγράφονται στην Παράγραφο 3.4. Ο καθορισμός των παραμέτρων αυτών χειροκίνητα καθίσταται αρκετά δύσκολος λόγω του ότι οι κατάλληλες τιμές μπορεί να διαφέρουν κατά τη διάρκεια κάθε βήματος του αλγορίθμου. Έτσι, προκειμένου να προσεγγιστούν κατάλληλες τιμές για τις παραμέτρους από τον συγκεκριμένο αλγόριθμο, εφαρμόζεται επιλογή μοντέλου.

Πιο συγκεκριμένα βάσει της εργασίας [31], εφαρμόζεται η μέθοδος ”grid-search” όπου στην ουσία ορίζονται εκθετικά αυξανόμενες τιμές των παραμέτρων και στη συνέχεια κάθε ομάδα παραμέτρων

δοκιμάζεται με την μέθοδο του cross validation. Όσο αφορά το cross validation, χρησιμοποιείται το n-fold cross validation όπου το σύνολο των δεδομένων χωρίζεται σε n υποσύνολα. Από τα n υποσύνολα τα n-1 χρησιμοποιούνται ως σύνολα δεδομένων εκπαίδευσης ενώ αυτό που απομένει χρησιμοποιείται ως σύνολο δεδομένων δοκιμής. Η διαδικασία αυτή επαναλαμβάνεται n φορές ώστε να δοκιμαστεί η επίδοση κάθε συνόλου παραμέτρων χωριστά με αποτέλεσμα στο τέλος να επιλέγεται το καλύτερο εξ αυτών.



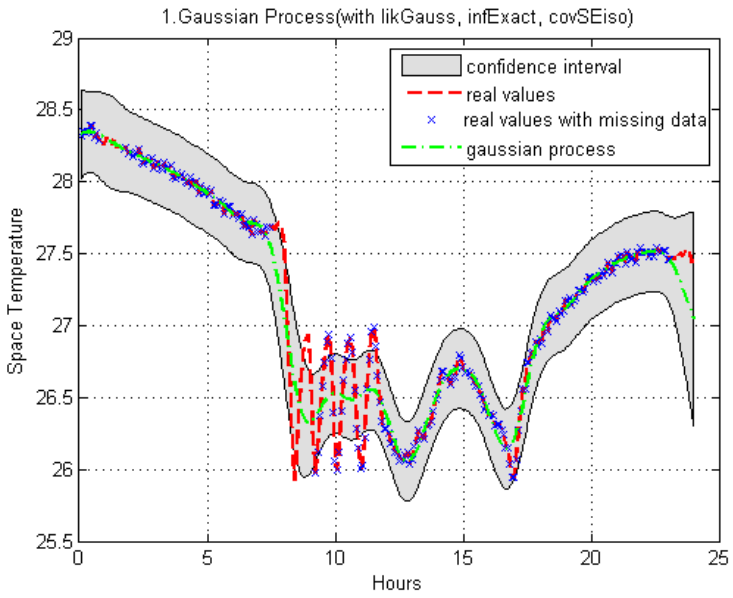
Σχήμα 4.9: Στην συγκεκριμένη γραφική παράσταση το πλήθος των δεδομένων που λείπουν είναι συγκεντρωμένο σε μία περιοχή.

Τα αποτελέσματα που παρουσιάζονται στα Σχήματα 4.8, 4.9 δείχνουν ότι η μέθοδος Support Vector Regression είναι σαφώς πιο σταθερή και σίγουρα πιο αξιόπιστη λόγω της λεπτομερής αναζήτησης για τη χρήση των σωστών παραμέτρων που αναφέρθηκε παραπάνω. Ακόμη παρατηρείται ότι όταν το κομμάτι των δεδομένων που λείπουν είναι μεγάλο σε σχέση με το συνολικό δείγμα των δεδομένων και τα δεδομένα δεν είναι ομαλά, όπως συμβαίνει στο Σχήμα 4.9, η αποτελεσματικότητα της μεθόδου μειώνεται. Αυτό συμβαίνει γιατί το μοντέλο εκτίμησης που παράγεται εξαρτάται μόνο από ένα υποσύνολο δεδομένων εκπαίδευσης και όχι από το σύνολο τους, όπως περιγράφεται στην Παράγραφο 3.4.

Gaussian Processes

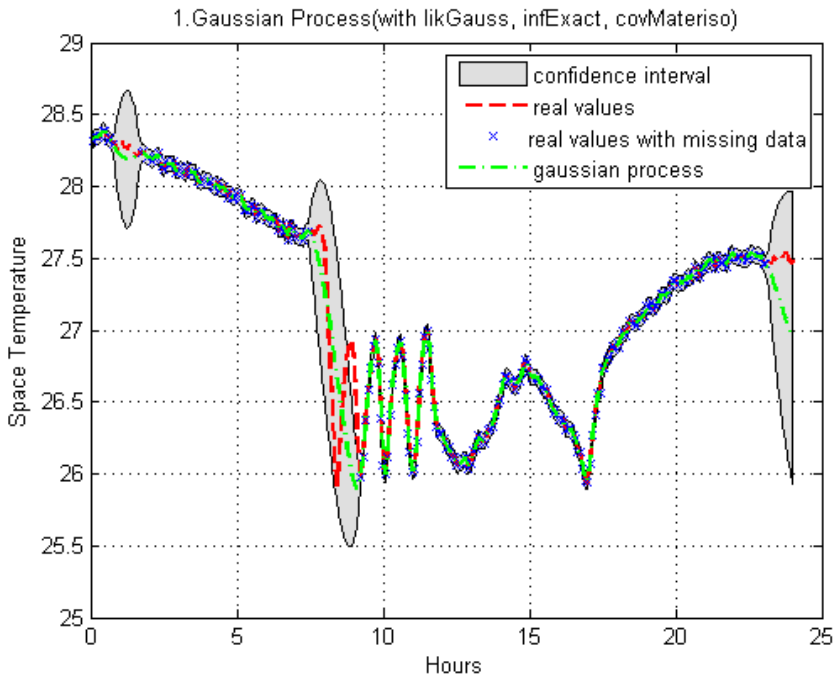
Κύριο χαρακτηριστικό αυτής της μεθόδου είναι η δυνατότητα επιλογής της κατάλληλης covariance function ανάλογα με τα χαρακτηριστικά των δεδομένων που μελετούνται. Πιο συγκεκριμένα

στο Σχήμα 4.10 όπου χρησιμοποιείται η squared exponential συνάρτηση προκύπτει μια πιο ομαλή εκτίμηση της μεθόδου, στις περιοχές όπου λείπουν δεδομένα. Αυτό συμβαίνει διότι όπως έχει περιγραφτεί και στην Παράγραφο 3.5 η squared exponential συνάρτηση είναι απείρως παραγωγίσιμη άρα το αποτέλεσμα της Gaussian Process είναι πιο ομαλό. Ένα ακόμα στοιχείο που ενισχύει την παραπάνω παρατήρηση είναι το διάστημα εμπιστοσύνης (confidence interval) που έχει σχεδιαστεί στα σχήματα και το οποίο υπολογίζεται ως $m-2\sqrt{s}$ και $m+2\sqrt{s}$, με m μέση τιμή των προβλεπόμενων τιμών των δεδομένων και s την διακύμανση των προβλεπόμενων τιμών των δεδομένων. Στην προκειμένη περίπτωση το confidence interval είναι αρκετά ευρύ, δείχνοντας ότι το μοντέλο με την χρήση της squared exponential συνάρτησης συνδιακύμανσης δεν είναι τόσο ”βέβαιο” για τα συγκεκριμένα δεδομένα.



Σχήμα 4.10: Το συγκεκριμένο Gaussian process εκτελέστηκε χρησιμοποιώντας για covariance function την συνάρτηση covSEiso (βλέπε Παράγραφο 3.5).

Αντίθετα με τα προηγούμενα, στο Σχήμα 4.11 όπου για την εκτίμηση των δεδομένων από το Gaussian Process χρησιμοποιείται ως covariance function η Matern συνάρτηση, παρατηρείται καλύτερο αποτέλεσμα. Αυτό συμβαίνει καθώς υπάρχει η δυνατότητα ρύθμισης της παραμέτρου d της συγκεκριμένης συνάρτησης όπως περιγράφεται στην Παράγραφο 3.5, και κατ’ επέκταση ρύθμιση της ομαλότητας της εκτίμησης που προκύπτει από το Gaussian Process. Έτσι λόγω της μη ομαλότητας των δεδομένων στο συγκεκριμένο πείραμα, επιλέγοντας $d = 3$ η εκτίμηση των δεδομένων γίνεται με μεγάλη ακρίβεια κάτι που ενισχύεται και από το confidence interval που είναι πολύ μικρό αποδεικνύοντας την πολύ μικρή απόκλιση που μπορεί να έχει η εκτίμηση του Gaussian Process που χρησιμοποιεί Matern covariance function, από τα πραγματικά δεδομένα στο συγκεκριμένο πείραμα.



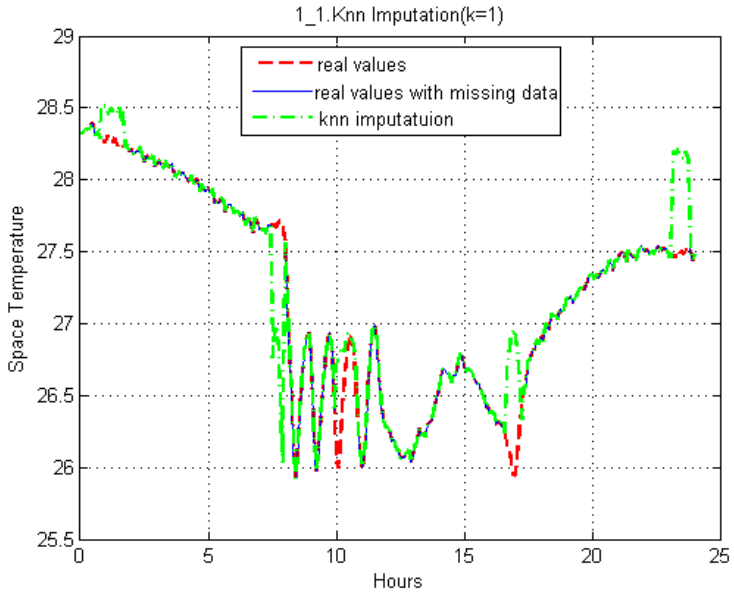
Σχήμα 4.11: Σε αυτό το πείραμα χρησιμοποιείται για covariance function η συνάρτηση covMaterniso με παράμετρο 3 (βλέπε Παράγραφο 3.5) με σαφώς καλύτερα αποτελέσματα σε σχέση με το Σχήμα 4.10.

Nearest Neighbor imputation

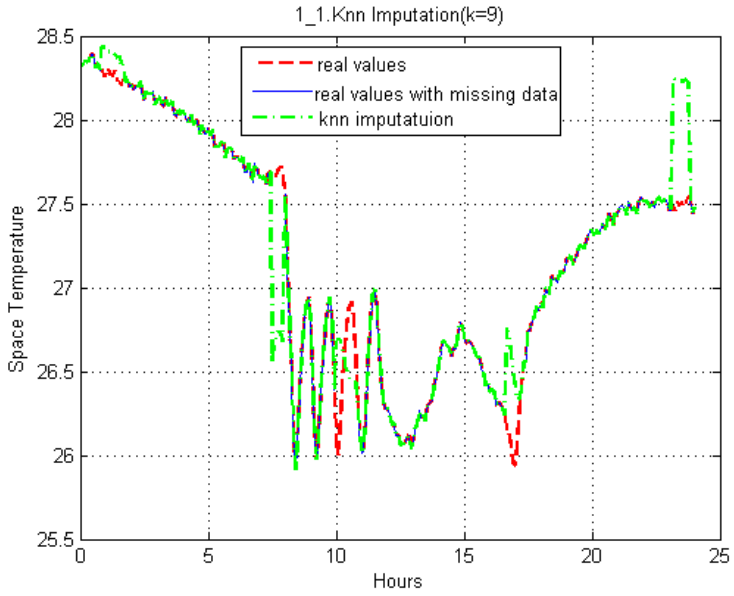
Η Nearest Neighbor imputation μέθοδος είναι η πιο απλή από όσες χρησιμοποιήθηκαν στην παρούσα διπλωματική εργασία. Στην συγκεκριμένη μέθοδο η διαφοροποίηση που μπορεί να γίνει όσο αφορά τον υπολογισμό της εκτίμησης των δεδομένων που λείπουν, είναι η επιλογή του πλήθους των πλησιέστερων γειτόνων.

Έτσι λοιπόν στα δύο επόμενα σχήματα, δηλαδή τα Σχήματα 4.12 και 4.13, παρουσιάζεται το αποτέλεσμα της εκτίμησης της μεθόδου για τον πλέον κοντινό γείτονα στο πρώτο σχήμα και για τους εννιά πιο κοντινούς γείτονες στο δεύτερο σχήμα. Αυτό που γίνεται εύκολα αντιληπτό είναι ότι δεν μπορεί να παρατηρηθεί διαφορετικό αποτέλεσμα στα δύο αυτά σχήματα διότι η διαφορά είναι μηδαμινή. Αυτή η πολύ μικρή διαφορά είναι ευδιάκριτη μόνο παρατηρώντας τις τιμές των δεικτών αποτελεσματικότητας των μεθόδων στον Πίνακα 4.2 και στον Πίνακα 4.4.

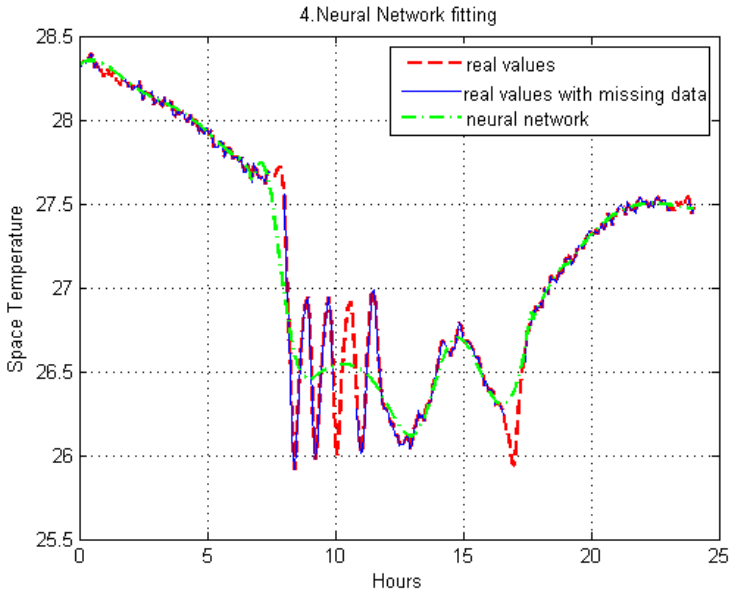
Η ομοιότητα του αποτελέσματος είτε χρησιμοποιώντας την τιμή της παραμέτρου $k = 1$ είτε $k = 9$, εξηγείται από το γεγονός ότι τα βάρη που δίνονται στους πλησιέστερους γείτονες είναι αντιστρόφως ανάλογα με την απόσταση τους από τα δεδομένα για τα οποία εκτελείται η μέθοδος. Έτσι με αυτό τον τρόπο ο πλησιέστερος γείτονας θα έχει πάντα την μεγαλύτερη επίδραση στο τελικό αποτέλεσμα. Το νόημα χρήσης περισσοτέρων του ενός πλησιέστερων γειτόνων βρίσκεται στην προσπάθεια εκμετάλλευσης έστω και μίας μικρής επιπλέον πληροφορίας που μπορούν να προσδώσουν οι περισσότεροι πλησιέστεροι γείτονες.



Σχήμα 4.12: Στο συγκεκριμένο πείραμα χρησιμοποιείται k=1, δηλαδή η μέθοδος στηρίζεται στον πλέον κοντινό γείτονα.



Σχήμα 4.13: Στο συγκεκριμένο πείραμα χρησιμοποιείται k=9, δηλαδή η μέθοδος στηρίζεται στους 9 πιο κοντινούς γείτονες.



Σχήμα 4.14: Data correction με χρήση Neural Network fitting.

Neural Network fitting

Στο Σχήμα 4.14 παρουσιάζεται το αποτέλεσμα της χρήσης της μεθόδου Neural Network fitting όπου η εκπαίδευση των νευρώνων γίνεται βάσει του αλγορίθμου Levenberg-Marquardt. Σε αυτή την περίπτωση γίνεται εύκολα αντιληπτό πως η συγκεκριμένη μέθοδος αποδίδει πολύ καλά σε περιοχές του δείγματος όπου τα δεδομένα είναι πιο ομαλά ενώ αντίθετα στην περιοχή που υπάρχει έντονη αυξομείωση των τιμών η μέθοδος δεν αποδίδει.

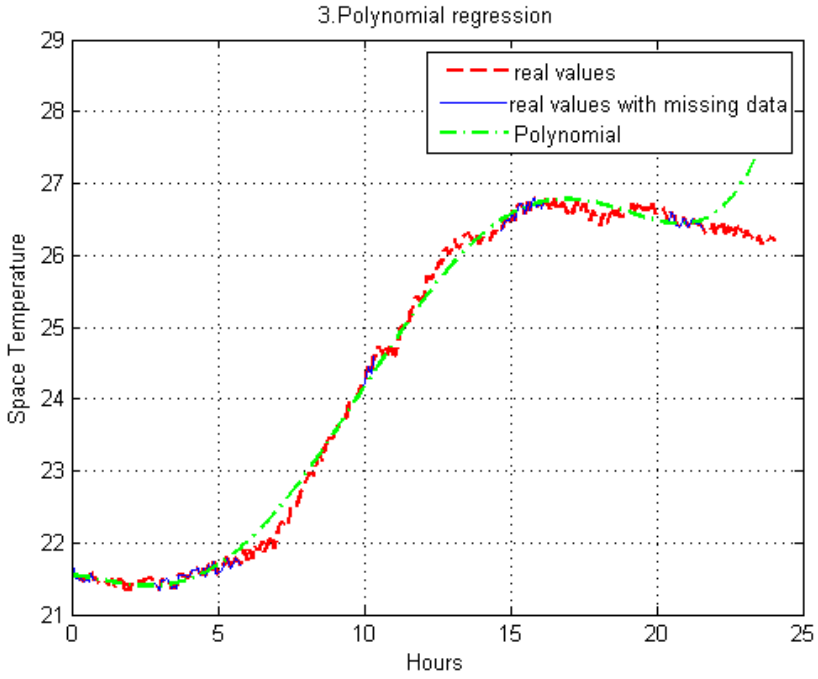
4.3 Πείραμα 2

Στο Πείραμα 2 χρησιμοποιήθηκαν δεδομένα μίας ολόκληρης ημέρας αλλά αυτή τη φορά τα δεδομένα παρουσιάζουν μία πολύ ομαλή μορφή, σε αντίθεση με το Πείραμα 1 όπου οι τιμές των δεδομένων είχαν έντονες αυξομειώσεις. Πιο συγκεκριμένα, στις γραφικές παραστάσεις που ακολουθούν, οι μέθοδοι εφαρμόστηκαν χρησιμοποιώντας ένα δείγμα θερμοκρασιών του γραφείου μίας ολόκληρης μέρας όπου λείπουν δεδομένα περίπου 16 ωρών.

Όσο αφορά τους συμβολισμούς και τους χρωματισμούς των γραφικών παραστάσεων διατηρήθηκαν οι ίδιοι όπως ακριβώς χρησιμοποιήθηκαν και στο Πείραμα 1.

Polynomial Regression

Στο Σχήμα 4.15 όπου παρουσιάζεται η μέθοδος του Polynomial Regression, γίνεται άμεσα αντιληπτό το πολύ καλό αποτέλεσμα της μεθόδου όσο αφορά την εκτίμηση των δεδομένων. Αυτό οφείλεται στο γεγονός ότι τα δεδομένα του συγκεκριμένου πειράματος είναι πολύ ομαλά και αρκετά γραμμικά. Έτσι χρησιμοποιώντας και σε αυτή την περίπτωση πολυώνυμο $5^{ου}$ η επίδοση της μεθόδου είναι υψηλή



Σχήμα 4.15: Στην παραπάνω γραφική παράσταση το πολυώνυμο που χρησιμοποιείται είναι 5ου βαθμού και το regression παρουσιάζει καλό αποτέλεσμα λόγω της γραμμικότητας των δεδομένων.

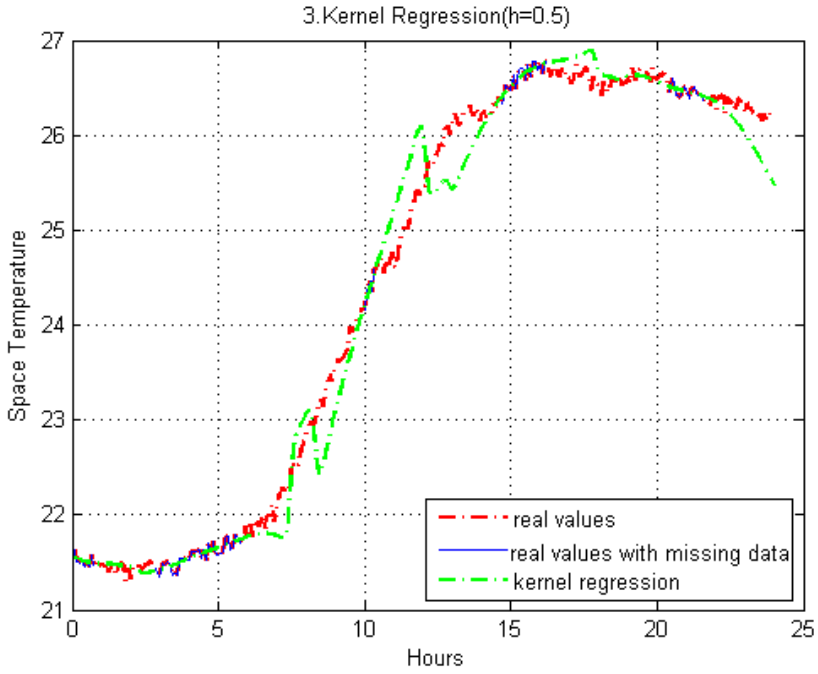
παρόλο που το πλήθος των δεδομένων που λείπουν είναι αρκετά μεγάλο, κάτι όμως που λόγω της γραμμικότητας των δεδομένων δεν επηρεάζει την συγκεκριμένη μέθοδο διότι της αρκούν λιγότερα δεδομένα εκπαίδευσης για να ολοκληρώσει την εκτίμηση των δεδομένων που λείπουν.

Local(Kernel) Regression

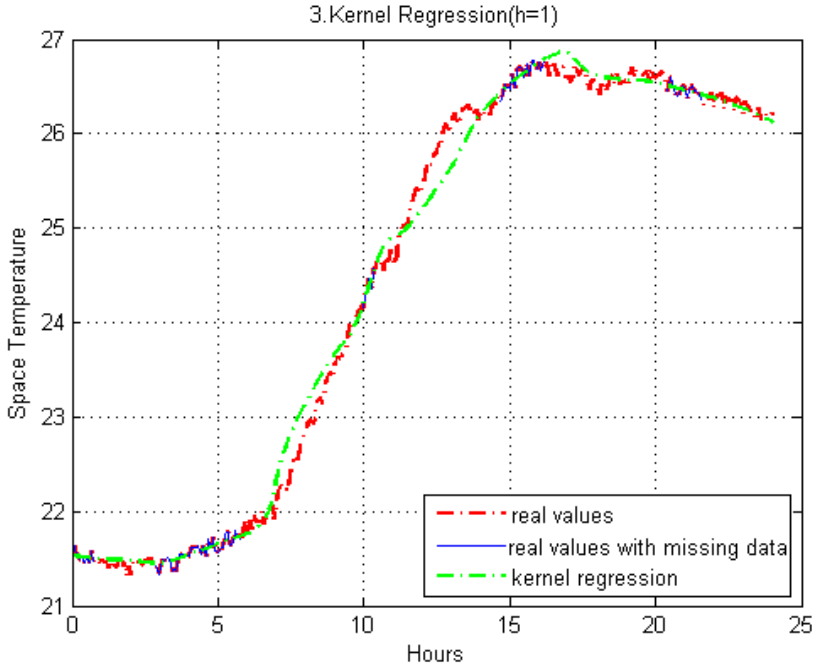
Κινούμενοι προς την ίδια κατεύθυνση με το Πείραμα 1 όσο αφορά την μέθοδο του Local Regression, στα Σχήματα 4.16, 4.17 και 4.18 παρουσιάζονται τα αποτελέσματα της μεθόδου εφαρμόζοντας στην παράμετρο h του αλγορίθμου τις τιμές $0.5,\ 1$ και 1.7 για κάθε σχήμα αντίστοιχα.

Στο συγκεκριμένο πείραμα παρατηρείται ότι όταν χρησιμοποιείται μικρή τιμή στην παράμετρο h (Σχήμα 4.16), ο αλγόριθμος προσπαθεί να αποδώσει λεπτομερής εκτίμηση των δεδομένων όμως αποτυγχάνει καθώς λείπουν πολλά δεδομένα. Αυτό συμβαίνει διότι χρησιμοποιώντας μικρή τιμή στην παράμετρο h, η εκτίμηση των δεδομένων σε ένα σημείο γίνεται χρησιμοποιώντας δεδομένα εκπαίδευσης που βρίσκονται σε κοντινή περιοχή με αποτέλεσμα όταν λείπουν αρκετά δεδομένα η εκτίμηση να γίνεται με μη επαρκή δεδομένα εκπαίδευσης και άρα το τελικό αποτέλεσμα να μην είναι ικανοποιητικό. Η κατάσταση αυτή διορθώνεται χρησιμοποιώντας μεγαλύτερες τιμές της παραμέτρου h δημιουργώντας πιο γενικευμένες εκτιμήσεις σαφώς αποτελεσματικότερες, όπως στα Σχήματα 4.17 και 4.18.

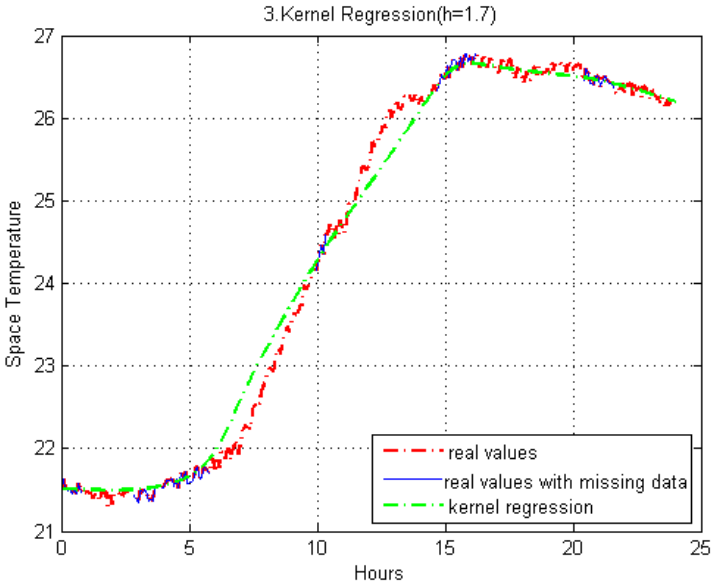
Στα Σχήματα 4.19 και 4.20 παρακάτω παρουσιάζεται η διαφοροποίηση του αποτελέσματος όσο αφορά την βέλτιστη τιμή της παραμέτρου h όταν διαφέρει το πλήθος των δεδομένων που λείπουν για το ίδιο δείγμα δεδομένων.



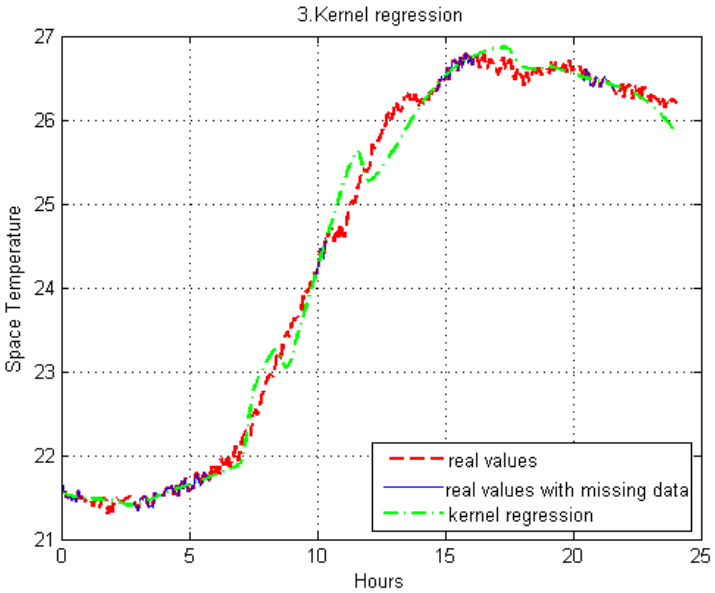
Σχήμα 4.16: Επιλογή της παραμέτρου εξομάλυνσης h=0.5



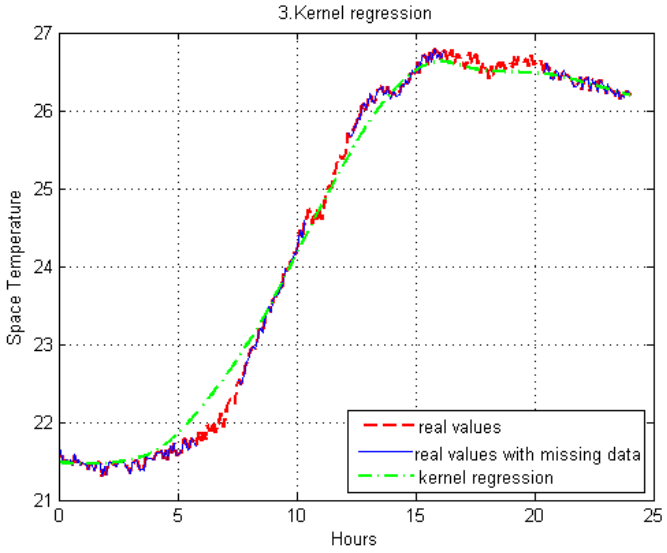
Σχήμα 4.17: Επιλογή της παραμέτρου εξομάλυνσης h=1



Σχήμα 4.18: Επιλογή της παραμέτρου εξομάλυνσης h=1.7



Σχήμα 4.19: Χρήση του τρόπου που περιγράφεται στην Παράγραφο 3.3 για τον υπολογισμό της παραμέτρου εξομάλυνσης. Σε αυτή την περίπτωση προκύπτει h=0.6809



Σχήμα 4.20: Χρήση του τρόπου που περιγράφεται στην Παράγραφο 3.3 για τον υπολογισμό της παραμέτρου εξομάλυνσης. Σε αυτή την περίπτωση τα missing data είναι σαφώς λιγότερα από το Σχήμα 4.19 οπότε προκύπτει h=1.8351

Support Vector Regression

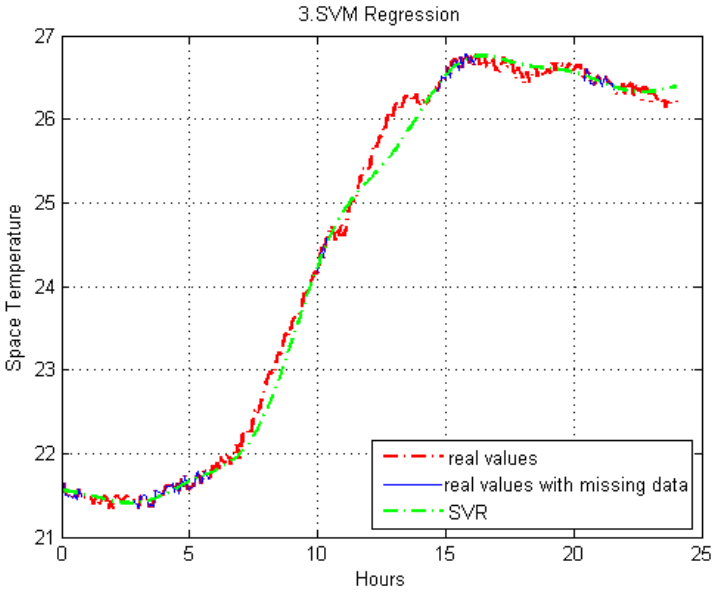
Το αποτέλεσμα της μεθόδου Support Vector Regression που παρουσιάζεται στο Σχήμα 4.21 είναι αρκετά ικανοποιητικό παρά τα μεγάλα διαστήματα των δεδομένων που λείπουν. Αυτό οφείλεται στο γεγονός πως παρόλο που στο συγκεκριμένο πείραμα λείπουν αρκετά δεδομένα, είναι ωστόσο αρκετά ομαλά παρουσιάζουν στην ουσία μία γραμμική αναπαράσταση άρα δεν είναι πολλά δεδομένα εκπαίδευσης συγκεντρωμένα σε μικρή περιοχή όπως συμβαίνει στο Πείραμα 1 με αποτέλεσμα ο αλγόριθμος να είναι πιο ”σίγουρος” για το αποτέλεσμα της εκτίμησης.

Gaussian Processes

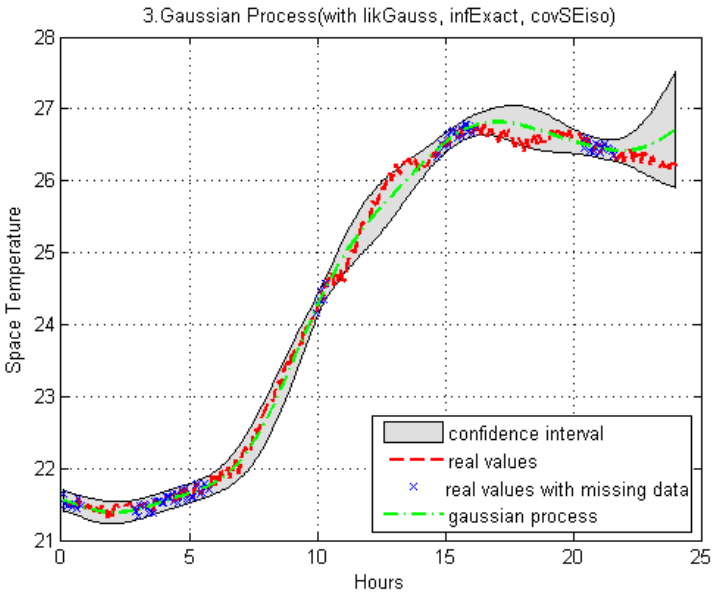
Στο Σχήμα 4.22 παρουσιάζεται το αποτέλεσμα της μεθόδου Gaussian Process χρησιμοποιώντας τα δεδομένα του Πειράματος 2. Όπως πολλάκις έχει αναφερθεί μέχρι αυτό το σημείο, τα δεδομένα του συγκεκριμένου πειράματος είναι αρκετά ομαλά οπότε στην συγκεκριμένη μέθοδο αρκεί να χρησιμοποιηθεί ως συνάρτηση συνδιακύμανσης μόνο η συνάρτηση Squared Exponential καθώς όπως αναφέρεται στην Παράγραφο 3.5 είναι απείρως παραγωγίσιμη με αποτέλεσμα να προσφέρει αρκετά ομαλή εκτίμηση μέσω της μεθόδου Gaussian Process. Επίσης παρατηρείται ότι το πλάτος του confidence interval είναι αρκετά μικρό κάτι που δείχνει την ”σιγουριά” της μεθόδου για την εκτίμηση των συγκεκριμένων δεδομένων.

Nearest Neighbor imputation

Τα επόμενα δύο σχήματα που ακολουθούν (4.23, 4.24) αφορούν την μέθοδο Nearest Neighbor imputation. Στα συγκεκριμένα πειράματα δεν υπάρχει κάτι επιπλέον να προστεθεί από όσα έχουν

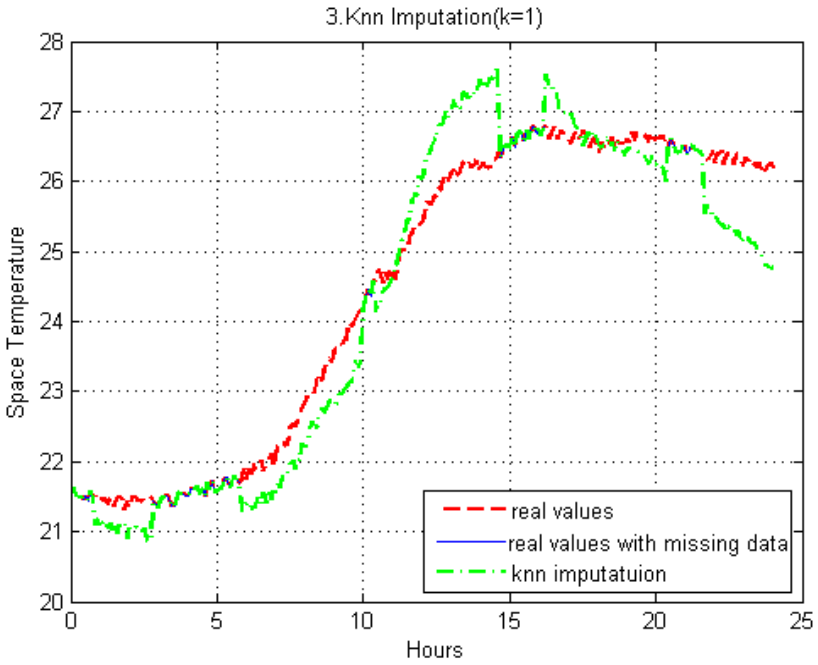


Σχήμα 4.21: Στην συγκεκριμένη γραφική παράσταση το πλήθος των δεδομένων που λείπουν είναι αρκετά μεγάλο σε σχέση με το μέγεθος του δείγματος που είναι ένα 24ωρο.



Σχήμα 4.22: Σε αυτή τη γραφική παράσταση φαίνεται η πολύ καλή λειτουργία του Gaussian process παρά το μεγάλο πλήθος των δεδομένων που λείπουν.

επισημανθεί στο προηγούμενο πείραμα, δηλαδή το Πείραμα 1. Η συγκεκριμένη μέθοδος συνεχίζει να έχει την ίδια συμπεριφορά με το Πείραμα 1 παρόλο που στο Πείραμα 2 τα δεδομένα είναι τελείως διαφορετικά.



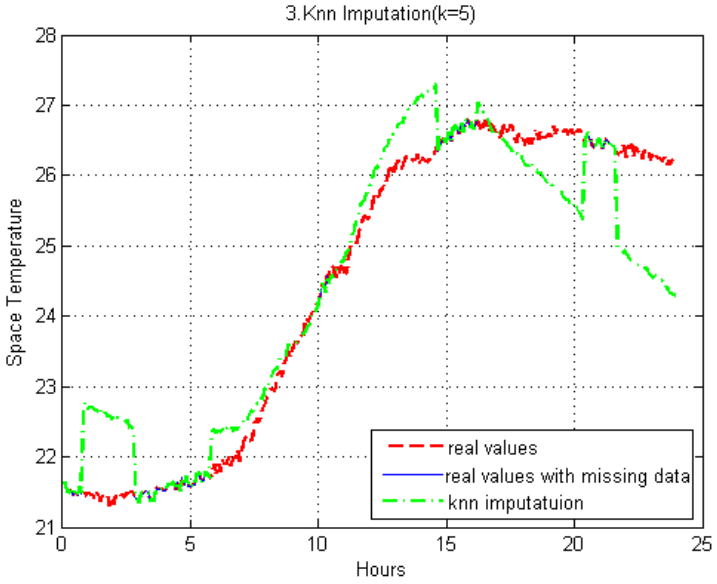
Σχήμα 4.23: Στο συγκεκριμένο πείραμα χρησιμοποιείται k=1, δηλαδή η μέθοδος στηρίζεται στον πλέον κοντινό γείτονα.

Neural Network fitting

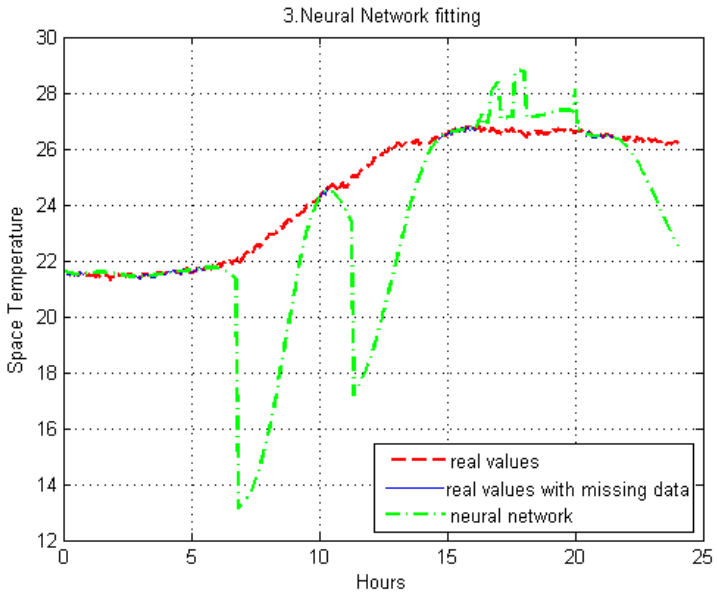
Στα Σχήματα 4.25, 4.26 παρουσιάζεται η διαφορά στην αποτελεσματικότητα της μεθόδου Neural Network fitting, όταν χρησιμοποιείται ως αλγόριθμος εκπαίδευσης ο Levenberg-Marquardt στο Σχήμα 4.25 και ο Bayesian regularization στο Σχήμα 4.26. Σε αυτά τα σχήματα παρουσιάζεται η επιρροή των μεγάλων διαστημάτων ελλειπόντων δεδομένων στη συγκεκριμένη μέθοδο.

Το αποτέλεσμα στο Σχήμα 4.25 οφείλετε στο γεγονός ότι ο αλγόριθμος Levenberg-Marquardt πρώτον επιδιώκει συνεχώς την ελαχιστοποίηση των σφαλμάτων και δεύτερον είναι εξαρτημένος από την αρχική υπόθεση για τις παραμέτρους του δικτύου. Έτσι ανάλογα με τα αρχικά βάρη του δικτύου, ο αλγόριθμος μπορεί να συγκλίνει σε ένα τοπικό ελάχιστο, όπως συμβαίνει και σε αυτή την περίπτωση.

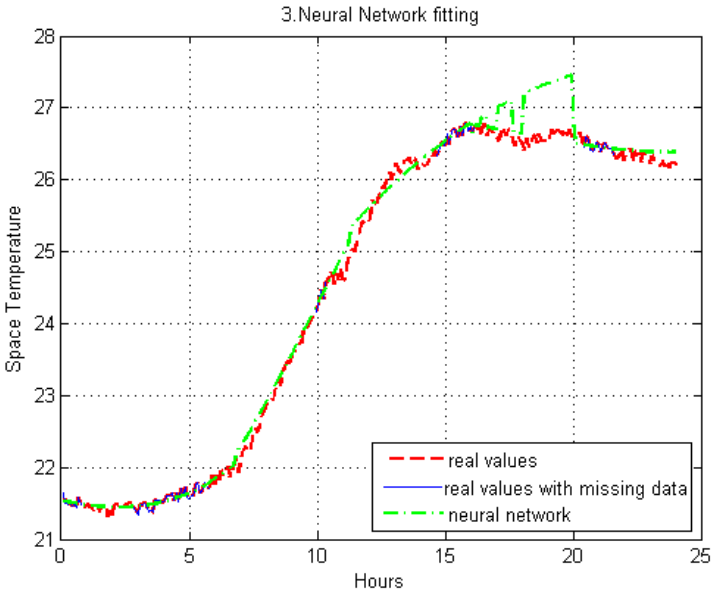
Αντίθετα ο αλγόριθμος Bayesian regularization, δεν εστιάζει μόνο στην ελαχιστοποίηση των σφαλμάτων αλλά σε ένα συνδυασμό ελαχιστοποίησης σφαλμάτων και βαρών του δικτύου, χρησιμοποιώντας παραμέτρους που κατευθύνουν την εκπαίδευση προς την ελαχιστοποίηση σφαλμάτων ή βαρών ανάλογα με το ποιο από τα δύο είναι το βέλτιστο σε κάθε επανάληψη του αλγορίθμου. Έτσι αποφεύγεται η δημιουργία τοπικών ελαχίστων όπως φαίνεται και στο Σχήμα 4.26, άρα ο αλγόριθμος είναι πιο αποδοτικός για τα δεδομένα του συγκεκριμένου Πειράματος.



Σχήμα 4.24: Στο συγκεκριμένο πείραμα χρησιμοποιείται k=5, δηλαδή η μέθοδος στηρίζεται στους 5 πιο κοντινούς γείτονες.



Σχήμα 4.25: Σε αυτό το νευρωνικό δίκτυο η εκπαίδευση πραγματοποιήθηκε με τον αλγόριθμο Levenberg-Marquardt.



Σχήμα 4.26: Σε αυτό το νευρωνικό δίκτυο η εκπαίδευση πραγματοποιήθηκε με τον αλγόριθμο Bayesian regularization.

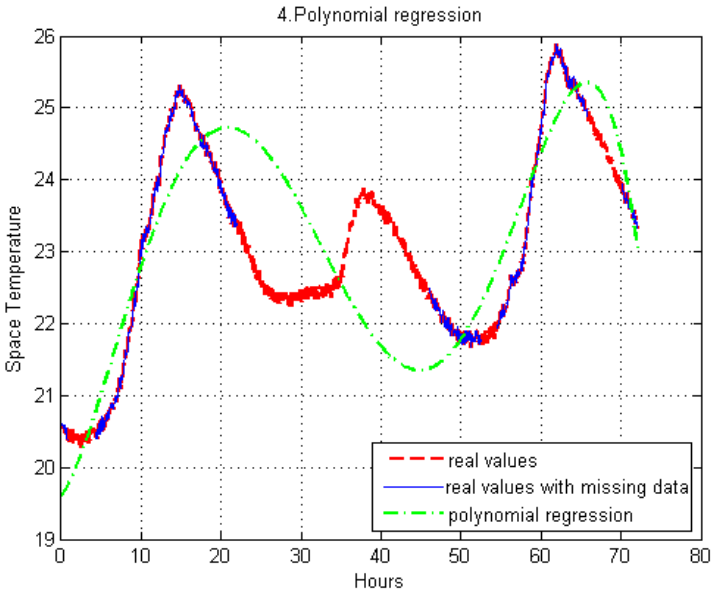
4.4 Πείραμα 3

Στο συγκεκριμένο πείραμα, το δείγμα των θερμοκρασιών του γραφείου αποτελείται συνολικά από τρεις ημέρες όπου λείπουν συνεχόμενα δεδομένα περίπου 24 ωρών. Η μεταβολή της θερμοκρασίας κατά την διάρκεια των τριών ημερών όσο αφορά τα πραγματικά δεδομένα, παρουσιάζει αυξομειώσεις με σταθερή συχνότητα άρα έχουμε μία σχετικά ομαλή γραφική αναπαράσταση. Όμοια με τα προηγούμενα πειράματα δεν έχει αλλάξει κάτι σχετικά με τους συμβολισμούς και τους χρωματισμούς των γραφικών παραστάσεων για τους οποίους ισχύει ότι με την κόκκινη διακεκομμένη γραμμή αναπαριστώνται τα πραγματικά δεδομένα ενώ με την μπλε γραμμή αναπαριστώνται τα ίδια δεδομένα αλλά με ελλείποντες τιμές. Τέλος με την πράσινη γραμμή παρουσιάζονται τα αποτελέσματα της κάθε μεθόδου.

Polynomial Regression

Στο Σχήμα 4.27 η μέθοδος Polynomial Regression τείνει να εκτιμήσει τα δεδομένα αλλά τελικά δεν το πετυχαίνει αποτελεσματικά. Αυτό οφείλεται στα χαρακτηριστικά των δεδομένων που χρησιμοποιούνται. Πιο συγκεκριμένα, αυτό που κάνει την εκτίμηση της μέθοδο να πλησιάζει τα πραγματικά δεδομένα είναι το γεγονός ότι τα δεδομένα του Πειράματος 3 παρουσιάζουν μία σχετική ομαλότητα, δεν έχουν δηλαδή έντονες αυξομειώσεις, όπως συμβαίνει και στα δεδομένα του Πειράματος 2. Από την άλλη μεριά το μεγάλο κενό δεδομένων μεταξύ της $22^{ης}$ και της $45^{ης}$ ώρας, μειώνει δραματικά το μέγεθος της χρήσιμης πληροφορίας που θα καθιστούσε τη μέθοδο αποτελεσματική.

Local(Kernel) Regression



Σχήμα 4.27: Στην παραπάνω γραφική παράσταση το πολυώνυμο που χρησιμοποιείται είναι 5ου βαθμού και το regression δεν παρουσιάζει καλό αποτέλεσμα λόγω της περιοδικότητας που φαίνονται να έχουν τα δεδομένα παρά την ομαλότητα τους.

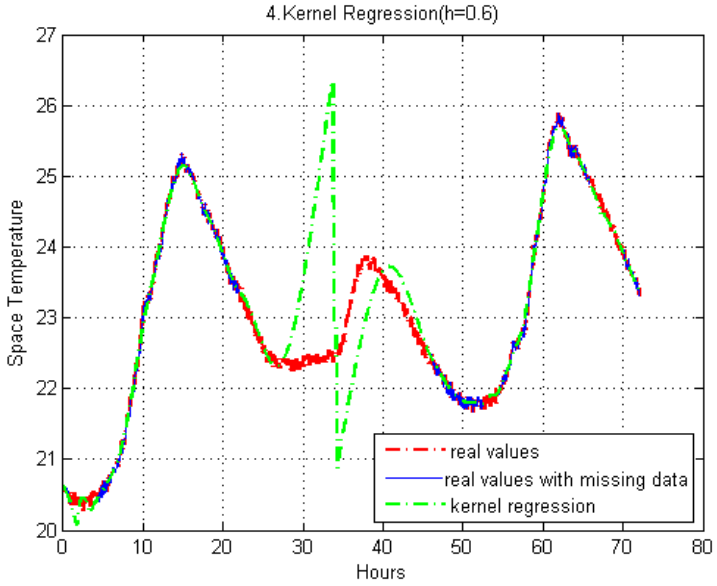
Στα Σχήματα 4.28, 4.29 και 4.30 κρίσιμος παράγοντας για την αποτελεσματικότητα της μεθόδου Local Regression αποτελεί η επιλογή της παραμέτρου εξομάλυνσης. Παρατηρώντας τα σχήματα γίνεται αντιληπτό ότι στο κομμάτι που υπάρχει μεγάλη έλλειψη δεδομένων υπάρχει όμοια αντίδραση της μεθόδου με αυτή στο Πείραμα 2. Δηλαδή όσο η τιμή της παραμέτρου h είναι μικρή, ναι μεν η εκτίμηση της μεθόδου είναι πολύ καλή στα διαστήματα όπου τα κενά είναι μικρά, αλλά στο διάστημα όπου λείπουν πολλά δεδομένα η μέθοδος δεν αποδίδει σύμφωνα με την ίδια λογική που ισχύει και στο Πείραμα 2, δηλαδή ότι όσο πιο μικρό είναι το h η μέθοδος χρησιμοποιεί τα πιο κοντινά δεδομένα ως δεδομένα εκπαίδευσης.

Έτσι λοιπόν και σε αυτή της περίπτωση του Πειράματος 3, αυξάνοντας την τιμή της παραμέτρου εξομάλυνσης παράγεται μια πιο γενικευμένη εκτίμηση στο πλήθος των δεδομένων που προσδίδει όμως καλύτερο συνολικό αποτέλεσμα σε σχέση με τα επιμέρους καλά αποτελέσματα, όταν το h είναι μικρό, στα σημεία που δεν λείπουν πολλά δεδομένα.

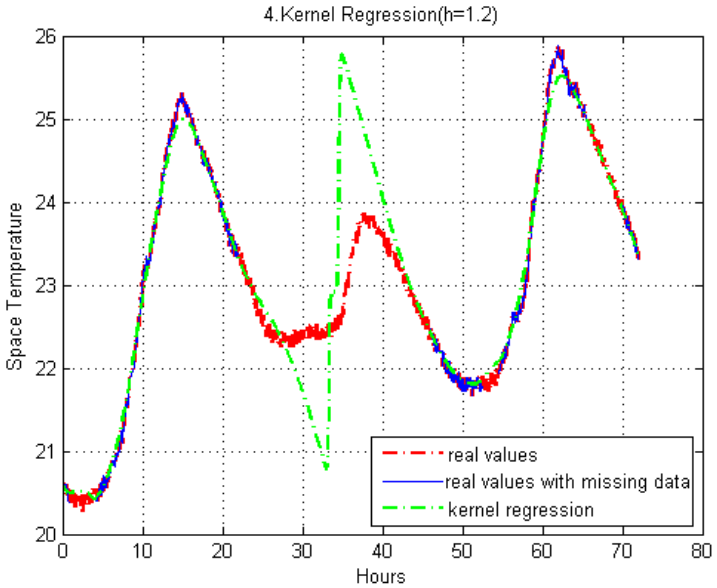
Όμοια στο Σχήμα 4.31 γίνεται χρήση του τρόπου υπολογισμού της βέλτιστης παραμέτρου εξομάλυνσης σύμφωνα με την Παράγραφο 3.3, αποδεικνύεται ότι στην συγκεκριμένη περίπτωση απαιτείται χρήση μεγάλης τιμής για την παράμετρο h ώστε να αποδώσει όσο το δυνατόν καλύτερη εκτίμηση η συγκεκριμένη μέθοδος.

Support Vector Regression

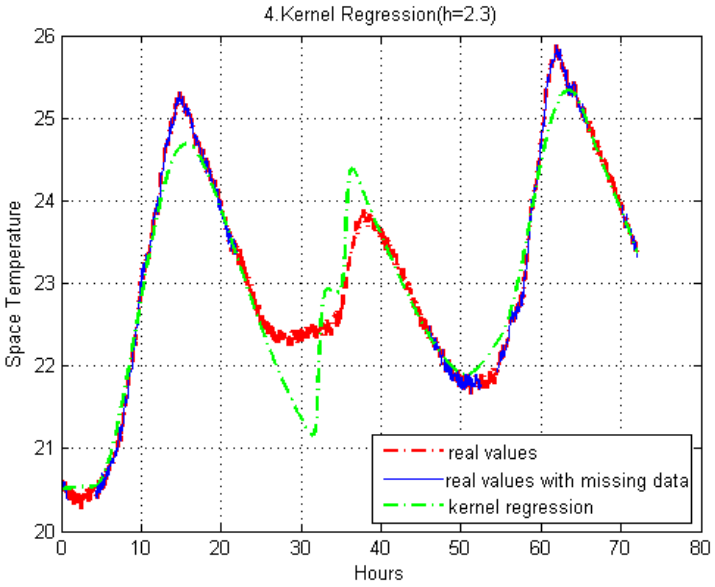
Όπως συμβαίνει στο Πείραμα 1 έτσι και σε αυτό το πείραμα που παρουσιάζεται στο Σχήμα 4.32 που αφορά τη μέθοδο Support Vector Regression, παρατηρείται η μη αποτελεσματική εκτίμηση στο διάστημα όπου υπάρχει μεγάλη έλλειψη δεδομένων σε αντίθεση με τα υπόλοιπα τμήματα της εκτί-



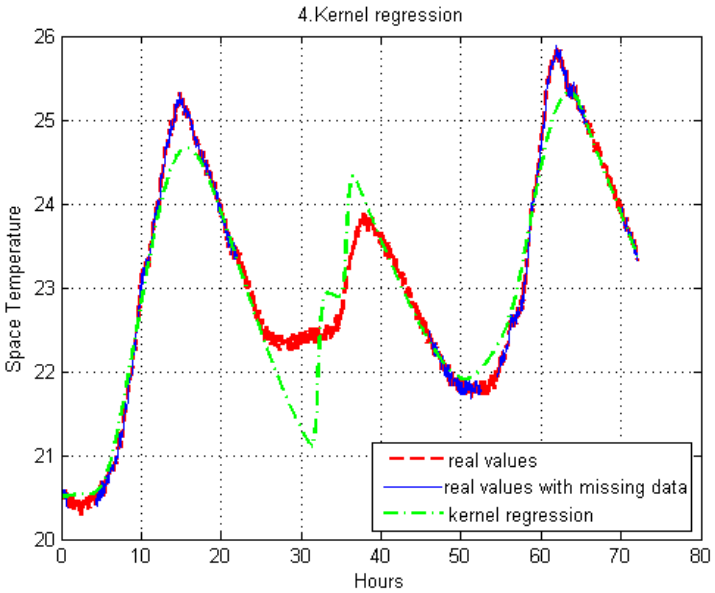
Σχήμα 4.28: Επιλογή της παραμέτρου εξομάλυνσης *h*=0.6



Σχήμα 4.29: Επιλογή της παραμέτρου εξομάλυνσης *h*=1.2

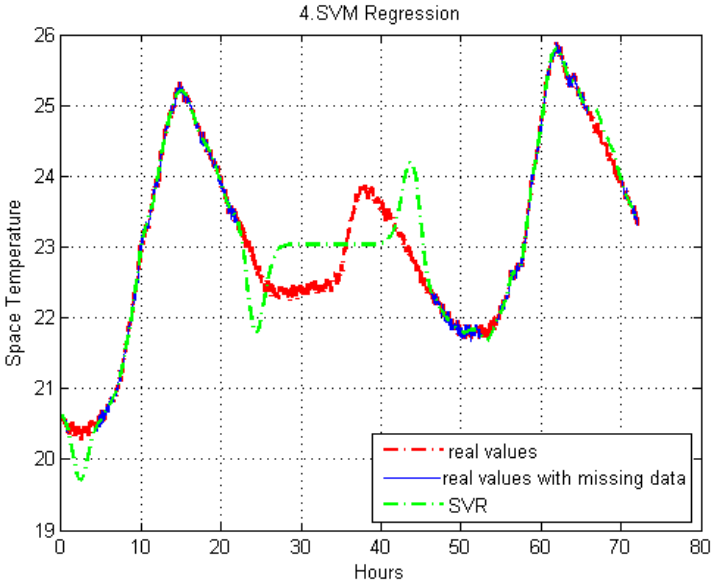


Σχήμα 4.30: Επιλογή της παραμέτρου εξομάλυνσης h=2.3



Σχήμα 4.31: Χρήση του τρόπου που περιγράφεται στην Παράγραφο 3.3 για τον υπολογισμό της παραμέτρου εξομάλυνσης. Σε αυτή την περίπτωση προκύπτει h=2.3893

μησης όπου η επίδοση της μεθόδου είναι πολύ καλή λόγω των μικρών διαστημάτων έλλειψης δεδομένων. Η συμπεριφορά της συγκεκριμένης μεθόδου όπως έχει προαναφερθεί οφείλεται στο γεγονός ότι η εκτίμηση σε ένα σημείο των δεδομένων στηρίζεται σε ένα υποσύνολο δεδομένων εκπαίδευσης πλησίον στο σημείο που γίνεται η εκτίμηση και όχι από το σύνολο των δεδομένων εκπαίδευσης.



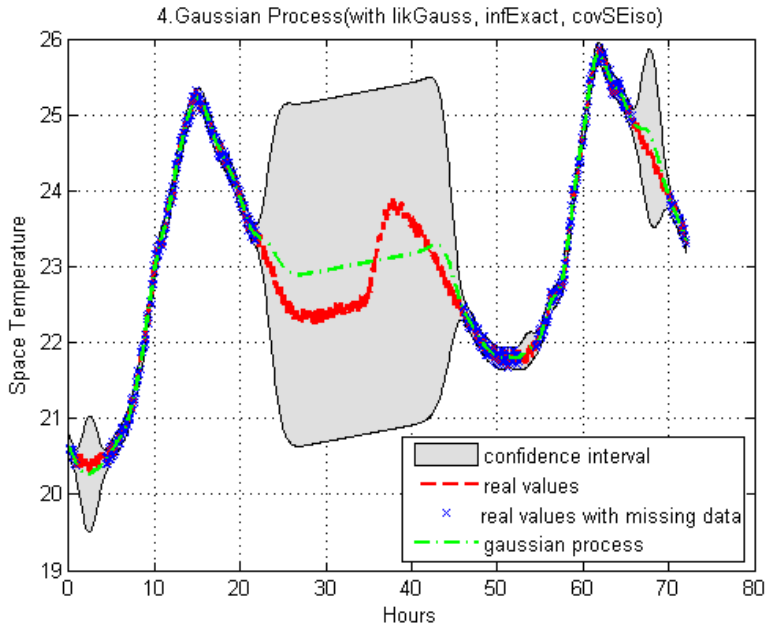
Σχήμα 4.32: Στην συγκεκριμένη γραφική παράσταση το πλήθος των δεδομένων που λείπουν είναι αρκετά μεγάλο.

Gaussian Processes

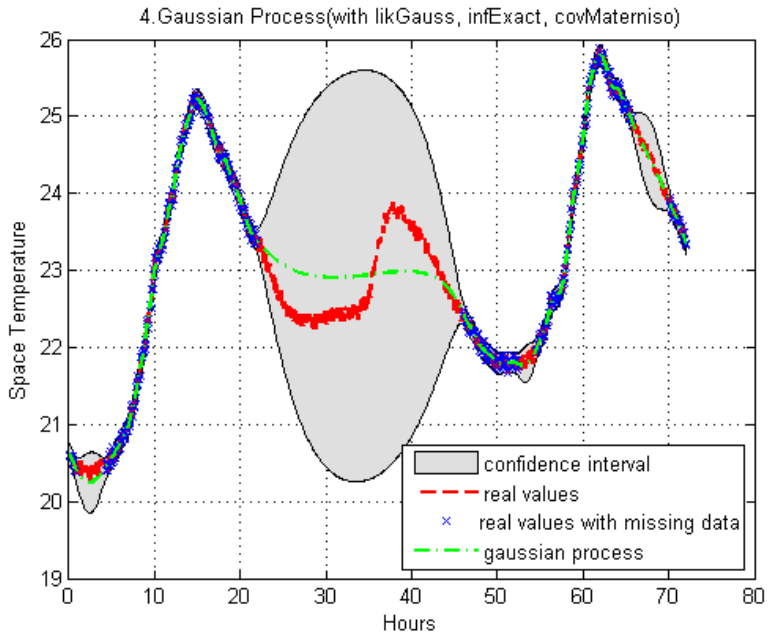
Στα επόμενα σχήματα (4.33, 4.34) που ακολουθούν και αναπαριστούν το αποτέλεσμα της Gaussian Process μεθόδου, παρατηρείται ότι και οι δύο προσεγγίσεις, στο Σχήμα 4.33 όπου χρησιμοποιείται η Squared Exponential συνάρτηση συνδιακύμανσης και στο Σχήμα 4.34 όπου χρησιμοποιείται η Matern συνάρτηση συνδιακύμανσης, η μέθοδος είναι αρκετά συντηρητική στο κομμάτι όπου λείπουν πολλά δεδομένα λόγω της μεγάλης αβεβαιότητας που προκύπτει όπως φαίνεται και από το confidence interval. Στα σημεία όπου τα δεδομένα που λείπουν είναι λιγότερα προβάδισμα έχει η Matern συνάρτηση συνδιακύμανσης που όπως έχει προαναφερθεί δίνει την δυνατότητα προσαρμογής της ομαλότητάς της μέσω της παραμέτρου d.

Nearest Neighbor imputation

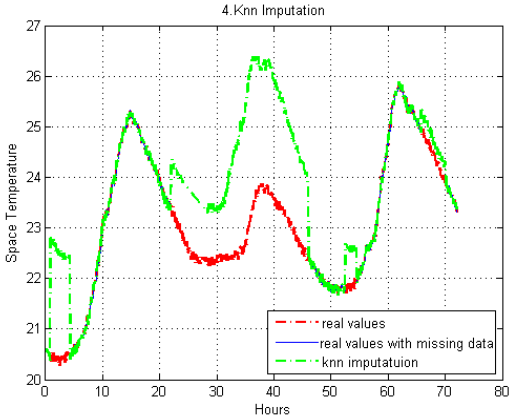
Η μέθοδος Nearest Neighbor imputation (Σχήμα 4.35) στο συγκεκριμένο πείραμα κρίνεται λιγότερο αναποτελεσματική από τα άλλα δύο πειράματα λόγω του μεγάλου μεγέθους του δείγματος των δεδομένων. Το κάθε δείγμα σε αυτό το πείραμα αντιπροσωπεύει συνολικά τρεις ημέρες οπότε είναι λογικό οι διαφορές στις συγκρίσεις των διαφόρων τριήμερων να είναι μεγαλύτερες απ' ότι μεταξύ μεμονωμένων ημερών.



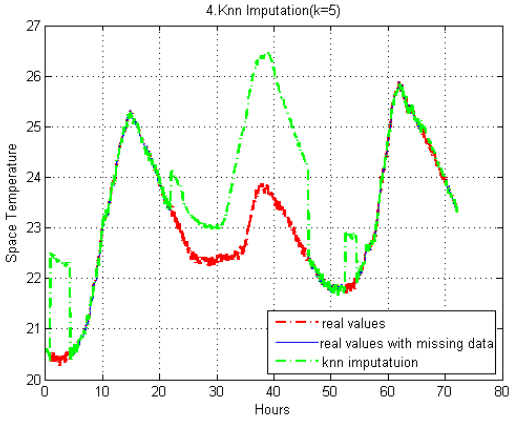
Σχήμα 4.33: Το συγκεκριμένο Gaussian process εκτελέστηκε χρησιμοποιώντας για covariance function την συνάρτηση covSEiso (βλέπε Παράγραφο 3.5).



Σχήμα 4.34: Σε αυτό το πείραμα χρησιμοποιείται για covariance function η συνάρτηση covMaterniso με παράμετρο 3 (βλέπε Παράγραφο 3.5). Αυτή τη φορά, μικρή βελτίωση παρατηρείται μόνο στο διάστημα σιγουριάς σε σχέση με το Σχήμα 4.33.



(α΄) Στο συγκεκριμένο πείραμα χρησιμοποιείται k=1, δηλαδή η μέθοδος στηρίζεται στον πλέον κοντινό γείτονα.

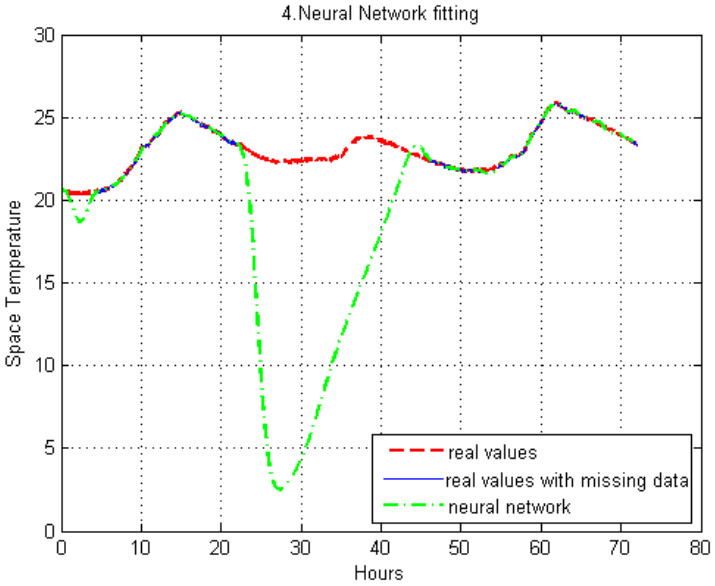


(β΄) Στο συγκεκριμένο πείραμα χρησιμοποιείται k=5, δηλαδή η μέθοδος στηρίζεται στους 5 πιο κοντινούς γείτονες.

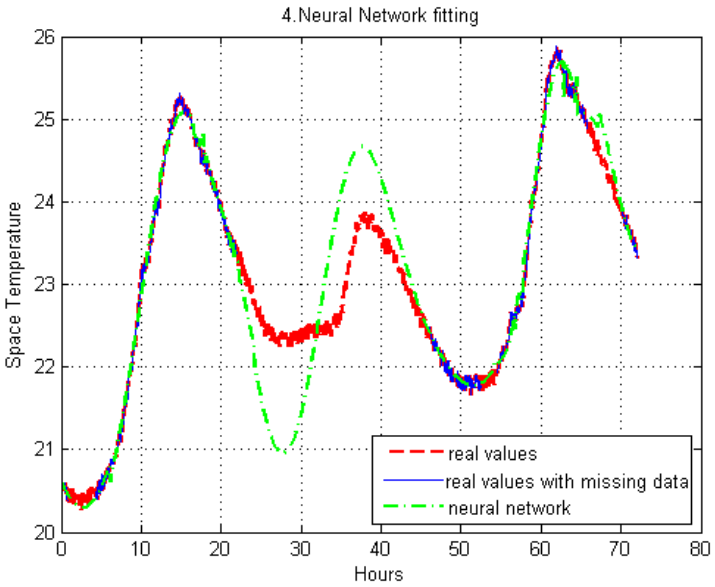
Σχήμα 4.35: Nearest Neighbor imputation για το Πείραμα 3.

Neural Network fitting

Στα Σχήματα 4.36, 4.37 παρουσιάζεται το αποτέλεσμα της μεθόδου Neural Network fitting για το δείγμα δεδομένων που χρησιμοποιείται στο Πείραμα 3. Στην προκειμένη περίπτωση και οι δύο αλγόριθμοι εκπαίδευσης των νευρώνων παρουσιάζουν τα ίδια αποτελέσματα όπως και με τα δεδομένα του Πειράματος 2. Δηλαδή ο μεν αλγόριθμος Levenberg-Marquardt δημιουργεί ένα έντονο τοπικό ελάχιστο στην περιοχή όπου λείπουν δεδομένα και ο δε Bayesian regularization είναι σαφώς πιο αποτελεσματικός αλλά σίγουρα πιο επηρεασμένος σε σχέση με το Πείραμα 2 από το μεγάλο κενό δεδομένων.



Σχήμα 4.36: Σε αυτό το νευρωνικό δίκτυο η εκπαίδευση πραγματοποιήθηκε με τον αλγόριθμο Levenberg-Marquardt.



Σχήμα 4.37: Σε αυτό το νευρωνικό δίκτυο η εκπαίδευση πραγματοποιήθηκε με τον αλγόριθμο Bayesian regularization.

Στη συνέχεια ακολουθούν συνοπτικοί πίνακες (Πίνακες 4.1, 4.2, 4.3, 4.4) των αποτελεσμάτων του συντελεστή προσδιορισμού R^2 (Παράγραφος 3.8.1) καθώς και για το μέσο τετραγωνικό σφάλμα (MSE) (Παράγραφος 3.8.2) για τα παραπάνω πειράματα. Οι τιμές που αναγράφονται στους παρακάτω πίνακες είναι ποσοστού επί τοις εκατό (%).

Πιο συγκεκριμένα, στους Πίνακες 4.1 και 4.2 παρουσιάζονται τα αποτελέσματα του συντελεστή προσδιορισμού R^2 που δείχνει πόσο κοντά είναι τα εκτιμώμενα δεδομένα στα πραγματικά δεδομένα. Στην ουσία οι συγκεκριμένοι πίνακες προσφέρουν μία πιο λεπτομερή απεικόνιση των αποτελεσμάτων των μεθόδων μέσω αριθμητικής περιγραφής, σε σχέση με τα σχήματα παραπάνω. Τα περιεχόμενα των πινάκων και οι γραφικές παραστάσεις των σχημάτων μελετήθηκαν από κοινού και για να παραχθούν τα συμπεράσματα συνδυάστηκαν οι πληροφορίες που προσφέρουν. Το κομμάτι στο οποίο συμβάλουν πιο πολύ οι πίνακες είναι η σύγκριση των αποτελεσμάτων της κάθε μεθόδου ξεχωριστά όταν τροποποιούνται οι παράμετροι, με σκοπό την κατανόηση της συμπεριφοράς της.

Με έντονους αριθμούς συμβολίζονται τα καλύτερα αποτελέσματα για το κάθε πείραμα. Όπου covSEiso και covMaterniso είναι οι συναρτήσεις συνδιακύμανσης Squared Exponential και Matern αντίστοιχα για την μέθοδο Gaussian Processes, LM και BR είναι οι αλγόριθμοι Levenberg-Marquardt και Bayesian regularization αντίστοιχα για την μέθοδο Neural Network fitting, k είναι το πλήθος των πλησιέστερων γειτόνων που χρησιμοποιούνται στη μέθοδο Nearest Neighbor imputation και h ο συντελεστής εξομάλυνσης της μεθόδου Local Regression. Τέλος όπου αναγράφεται neg σημαίνει πως το αποτέλεσμα της εκτίμησης είναι πολύ κακό με αποτέλεσμα η τιμή του R^2 να βγαίνει αρνητική.

Αντίστοιχα με τους δύο προηγούμενους πίνακες, οι Πίνακες 4.3 και 4.4 παρουσιάζουν το ποσοστό σφαλμάτων για τις μεθόδους. Αυτό που παρατηρείται στους συγκεκριμένους πίνακες είναι ότι όπου είναι τα καλύτερα αποτελέσματα για το R^2 στις αντίστοιχες θέσεις είναι τα καλύτερα αποτελέσματα για το MSE. Η μόνη διαφοροποίηση στο συμβολισμό είναι ότι όπου υπάρχει huge σημαίνει ότι το ποσοστό των σφαλμάτων είναι πολύ μεγάλο επιβεβαιώνοντας τους Πίνακες 4.1 και 4.2 για την μη αποτελεσματική λειτουργία των μεθόδων με τις αντίστοιχες παραμέτρους όπου εμφανίζεται αυτή η ένδειξη.

Πίνακας 4.1: Αποτελέσματα του συντελεστή προσδιορισμού R^2 για τα πειράματα των μεθόδων Polynomial Regression, Local Regression και Support Vector Regression.

	Polynomial Regression	Local Regression				Support Vector Regression
Πείραμα 1	87.22	h=0.1	h=0.5	h=1	h=0.97	94.74
		96.7	93.81	91.05	91.18	
Πείραμα 2	97.75	h=0.5	h=1	h=1.7	h=0.68	**99.47**
		98.42	99.4	99.14	99.17	
Πείραμα 3	46.95	h=0.6	h=1.2	h=1.7	h=2.39	91.86
		74.9	82.67	92.1	**93.06**	

Πίνακας 4.2: Αποτελέσματα του συντελεστή προσδιορισμού R^2 για τα πειράματα των μεθόδων Gaussian Processes, Neural Network fitting και Nearest Neighbor imputation.

	Gaussian Processes		Neural Network fitting		Nearest Neighbor imputation	
Πείραμα 1	covSEiso	covMaterniso	LM	BR	k=1	k=9
	94.65	**95.97**	-	94.62	88.38	90.46
Πείραμα 2	covSEiso	covMaterniso	LM	BR	k=1	k=9
	99.55	-	neg	98.69	92.93	88.89
Πείραμα 3	covSEiso	covMaterniso	LM	BR	k=1	k=9
	96.06	95.69	neg	93.46	25.13	28.5

Πίνακας 4.3: Αποτελέσματα του μέσου τετραγωνικού σφάλματος MSE για τα πειράματα των μεθόδων Polynomial Regression, Local Regression και Support Vector Regression.

	Polynomial Regression	Local Regression				Support Vector Regression
Πείραμα 1	6.61	h=0.1	h=0.5	h=1	h=0.97	2.72
		1.7	3.2	4.62	4.56	
Πείραμα 2	10.43	h=0.5	h=1	h=1.7	h=0.68	**2.45**
		7.29	2.77	3.97	3.86	
Πείραμα 3	98.37	h=0.6	h=1.2	h=1.7	h=2.39	15.1
		46.46	32.14	14.65	**12.87**	

Πίνακας 4.4: Αποτελέσματα του μέσου τετραγωνικού σφάλματος MSE για τα πειράματα των μεθόδων Gaussian Processes, Neural Network fitting και Nearest Neighbor imputation.

	Gaussian Processes		Neural Network fitting		Nearest Neighbor imputation	
Πείραμα 1	covSEiso	covMaterniso	LM	BR	k=1	k=9
	2.77	**2.05**	-	2.78	6.01	4.93
Πείραμα 2	covSEiso	covMaterniso	LM	BR	k=1	k=9
	2.08	-	huge	6.05	32.73	51.45
Πείραμα 3	covSEiso	covMaterniso	LM	BR	k=1	k=9
	7.3	8	huge	12.13	huge	huge

Κεφάλαιο 5

Πειράματα και Παράθεση Αποτελεσμάτων των Μεθόδων Πρόβλεψης Πληρότητας

Σε αυτό το κεφάλαιο, συνεχίζοντας με παρόμοια λογική όπως και στο προηγούμενο κεφάλαιο, παρατίθενται πειράματα που εκτελέστηκαν εφαρμόζοντας τις μεθόδους PreHeat και SmartThermostat που περιγράφονται στην Παράγραφο 3.10 και σχετίζονται με την πρόβλεψη πληρότητας κτηρίων χρησιμοποιώντας δεδομένα που παρέχουν πληροφορίες σχετικά με την παρουσία ενοίκων στα κτήρια.

5.1 Περιγραφή των Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν στα πειράματα πρόβλεψης πληρότητας των κτηρίων αντλήθηκαν από τις ίδιες πηγές δεδομένων από τις οποίες προέρχονται και τα δεδομένα που χρησιμοποιήθηκαν στα πειράματα των μεθόδων πρόβλεψης δεδομένων του Κεφαλαίου 4 και πιο συγκεκριμένα προέρχονται από το κτήριο της Cartif που περιγράφεται στη Παράγραφο 4.1.

Όπως στα δεδομένα που χρησιμοποιήθηκαν στις μεθόδους διόρθωσης δεδομένων έτσι και στα δεδομένα που χρησιμοποιήθηκαν στις μεθόδους πρόβλεψης πληρότητας, αρχικά πραγματοποιήθηκε interpolation με σκοπό να υπάρχει σταθερό χρονικό διάστημα πέντε λεπτών μεταξύ των τιμών λόγω του ότι όπως έχει προαναφερθεί η καταγραφή των δεδομένων γίνεται όταν υπάρχει μεταβολή των μετρούμενων τιμών κατά μήκος του χρόνου άρα όπως είναι φυσικό τα χρονικά διαστήματα μεταξύ των τιμών ποικίλουν. Επιπλέον, για να εξαντληθεί κάθε περιθώριο ακρίβειας στις μεθόδους του occupancy prediction, τα δεδομένα πληρότητας (occupancy) διαχωρίστηκαν όχι απλά σε καθημερινές μέρες και μέρες Σαββατοκύριακου, αλλά σε μεμονωμένες μέρες. Δηλαδή δημιουργήθηκαν πίνακες που αφορούν δεδομένα μόνο από Δευτέρες, μόνο από Τρίτες, κ.ο.κ.. Στη συνέχεια πραγματοποιήθηκε και ένας ακόμη διαχωρισμός μεταξύ ημερών που ανήκουν σε χειμερινούς μήνες και ημερών που ανήκουν σε καλοκαιρινούς μήνες. Οι παραπάνω διαχωρισμοί πραγματοποιήθηκαν διότι θεωρήθηκε αρκετά σημαντικό το γεγονός ότι η διαφοροποίηση από μέρα σε μέρα ή από εποχή σε εποχή στο ωράριο εργασίας των εργαζομένων του γραφείου του οποίου μελετήθηκαν τα δεδομένα, παίζει κύριο ρόλο στην πρόβλεψη πληρότητας του χώρου.

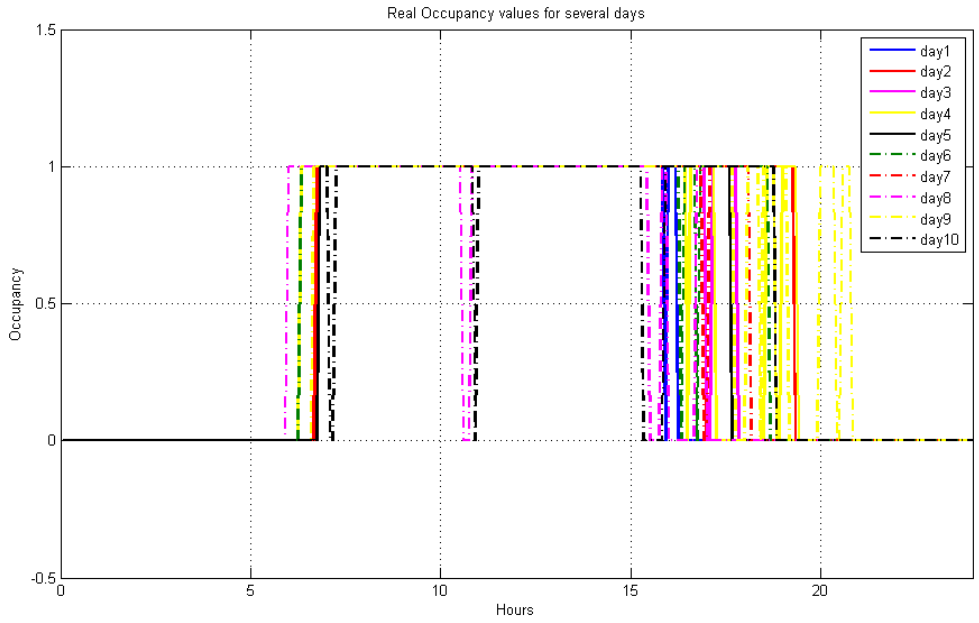
Ακόμη σχετικά με τα δεδομένα πληρότητας πρέπει να σημειωθεί ότι στο χώρο του γραφείου θεω-

ρείται ότι υπάρχει παρουσία ατόμων (occupied) όταν έστω και ένα άτομο βρίσκεται εντός γραφείου, τότε τα δεδομένα πληρότητας έχουν την τιμή ”1” ενώ στην αντίθετη περίπτωση έχουν την τιμή ”0”.

5.2 Πειράματα για τον SmartThermostat αλγόριθμο

Οι γραφικές παραστάσεις που παρουσιάζονται παρακάτω αφορούν πειράματα του αλγόριθμου SmartThermostat.Στα συγκεκριμένα πειράματα αναπαριστάται η πιθανότητα προβλεπόμενης πληρότητας του γραφείου, μέσα σε συγκεκριμένα διαστήματα δειγματοληψίας των 15 λεπτών, 30 λεπτών και μίας ώρας, για Δευτέρα, Τετάρτη, Πέμπτη, για μια οποιαδήποτε μέρα του χειμώνα ή οποιαδήποτε μέρα του καλοκαιριού. Πιο συγκεκριμένα η πιθανότητα αυτή υπολογίστηκε μέσω του μέσου όρου των τιμών του occupancy που περιλαμβάνονται σε κάθε διάστημα δειγματοληψίας. Αποτέλεσμα αυτού του υπολογισμού είναι, για κάθε διάστημα δειγματοληψίας να προκύπτει ποια είναι η πιθανότητα να θεωρηθεί ο χώρος occupied μέσα στο χρονικό διάστημα που ορίζει το διάστημα δειγματοληψίας.

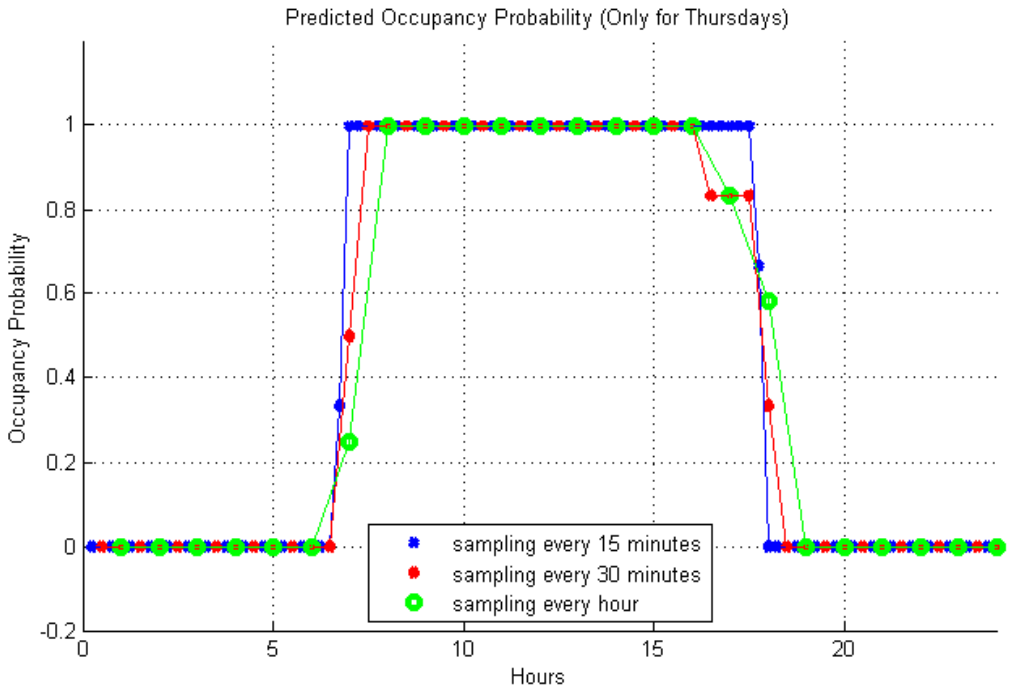
Στο Σχήμα 5.1 παρουσιάζονται 10 διαφορετικές μέρες που έχουν επιλεγεί τυχαία από το σύνολο των διαθέσιμων ημερών σε μία προσπάθεια κατανόησης της συμπεριφοράς των πραγματικών δεδομένων ούτως ώστε να γίνει καλύτερη αξιολόγηση των αποτελεσμάτων των προβλέψεων από τις μεθόδους που παρουσιάζονται παρακάτω.



Σχήμα 5.1: Στην παραπάνω γραφική παράσταση παρουσιάζονται πραγματικές τιμές του occupancy από 10 διαφορετικές μέρες.

Αρχικά παρουσιάζεται ένα πιο αναλυτικό διάγραμμα στο Σχήμα 5.2 ώστε να γίνει πιο κατανοητό το τι παρουσιάζουν τα διαγράμματα που ακολουθούν.

Πρώτα απ’ όλα το Σχήμα 5.2 απεικονίζει το αποτέλεσμα της πρόβλεψης του αλγορίθμου SmartThermostat θεωρώντας ότι η πρόβλεψη πραγματοποιείται για την ημέρα Πέμπτη και ως εκ τούτου στο

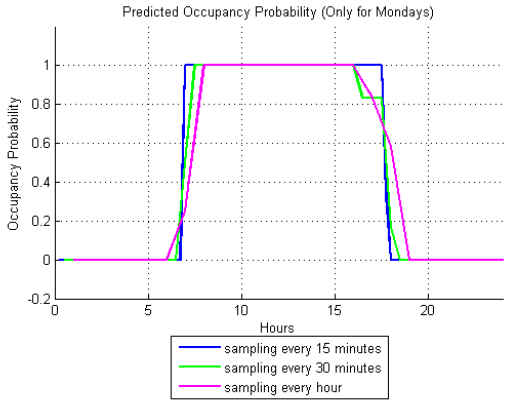


Σχήμα 5.2: Στην παραπάνω γραφική παράσταση παρουσιάζεται η πιθανότητα προβλεπόμενης πληρότητας του γραφείου για τη ημέρα Πέμπτη.

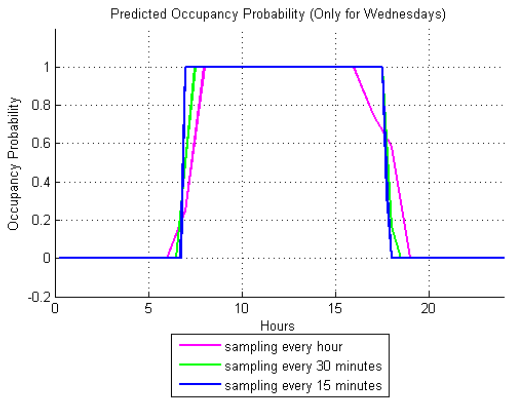
μοντέλο πρόβλεψης χρησιμοποιήθηκαν δεδομένα μόνο από Πέμπτες σύμφωνα με όσα περιγράφτηκαν στην Παράγραφο 4.1.2.

Παρατηρώντας το Σχήμα 5.2 αλλά και τα υπόλοιπα σχήματα των προβλέψεων, γίνεται αντιληπτό πως ο αλγόριθμος αδυνατεί να προβλέψει με ακρίβεια τις συχνές εναλλαγές του occupancy μετά τις 15:00 (βλέπε Σχήμα 5.1). Η μη πρόβλεψη των συχνών αυτών μεταβολών οφείλεται στο γεγονός ότι το χρονικό διάστημα που φεύγουν ή επανέρχονται στο γραφείο είναι πολύ σύντομο όπως επίσης και στο γεγονός ότι αυτές οι μεταβολές δεν έχουν σταθερό πρόγραμμα που συμβαίνουν, δηλαδή είναι τυχαίες και άρα διαφορετικές από μέρα σε μέρα.

Ένας τρόπος για να αντιμετωπιστεί αυτή η αδυναμία του αλγορίθμου να προβλέψει αυτό το κομμάτι των δεδομένων είναι ο υπολογισμός της πιθανότητας προβλεπόμενης πληρότητας του γραφείου που περιγράφτηκε στην Παράγραφο 5.2 προσπαθώντας έτσι να παρουσιαστεί μία πιο γενικευμένη προσέγγιση του προβλεπόμενου occupancy του γραφείου. Παρατηρώντας πιο προσεκτικά το Σχήμα 5.2 τα σημεία που φαίνονται και πιο συγκεκριμένα οι τελείες, μπλε για την δειγματοληψία ανά 15 λεπτά και κόκκινο για 30 λεπτά, όπως και οι πράσινοι κύκλοι για δειγματοληψία ανά 1 ώρα, ορίζουν τα διαστήματα των δειγματοληψιών κατά μήκος του 24ώρου, τα οποία έχουν μια συγκεκριμένη τιμή πιθανότητας προβλεπόμενου occupancy. Για παράδειγμα στο διάστημα μεταξύ 05:00 και 06:00 η πιθανότητα να προβλεφθεί occupancy είναι μηδέν. Στη συνέχεια για το διάστημα από τις 06:00 ως τις 07:00 η πιθανότητα να προβλεφθεί occupancy αυξάνεται περίπου στο 25%. Όμοια για το διάστημα από τις 17:00 ως τις 18:00 η πιθανότητα να προβλεφθεί occupancy μειώνεται από περίπου 82% στο 59%. Με τον ίδιο τρόπο εξηγείται το διάγραμμα και για τα υπόλοιπα διαστήματα δειγματοληψίας.



(α΄) Πιθανότητα προβλεπόμενης πληρότητας του γραφείου για τη ημέρα Δευτέρα



(β΄) Πιθανότητα προβλεπόμενης πληρότητας του γραφείου για τη ημέρα Τετάρτη

Σχήμα 5.3: Στις παραπάνω γραφικές παραστάσεις παρουσιάζεται η πιθανότητα προβλεπόμενης πληρότητας του γραφείου για δύο διαφορετικές ημέρες.

Στο Σχήμα 5.3 όπου ακολουθείτε η ιδία λογική με το Σχήμα 5.2, παρουσιάζεται η πρόβλεψη για την ημέρα Δευτέρα και την ημέρα Τετάρτη αντίστοιχα, χωρίς όμως να οδηγούμαστε σε διαφορετικά αποτελέσματα.

Παρόμοια κατάσταση συνεχίζεται και στο Σχήμα 5.4 όπου πραγματοποιήθηκε πρόβλεψη για την ανάδειξη τυχόν διαφοροποίησης στην τιμή του occupancy του γραφείου μεταξύ χειμερινών και καλοκαιρινών ημερών όπου λόγω των συνθηκών θα μπορούσε να υπάρχει και τροποποίηση του ωραρίου των εργαζομένων (άρα και διαφορετικές τιμές occupancy) πιθανότατα για την εκμετάλλευση του φυσικού φωτός ή την αποφυγή ακραίων θερμοκρασιών. Γι' αυτό το λόγο πραγματοποιήθηκε ο εν λόγω διαχωρισμός διότι δεν θα ήταν επιθυμητή η παρεμβολή δεδομένων από μια καλοκαιρινή ημέρα στην πρόβλεψη μιας χειμερινής ημέρας και αντίστροφα.

5.3 Πειράματα για τον Preheat αλγόριθμο

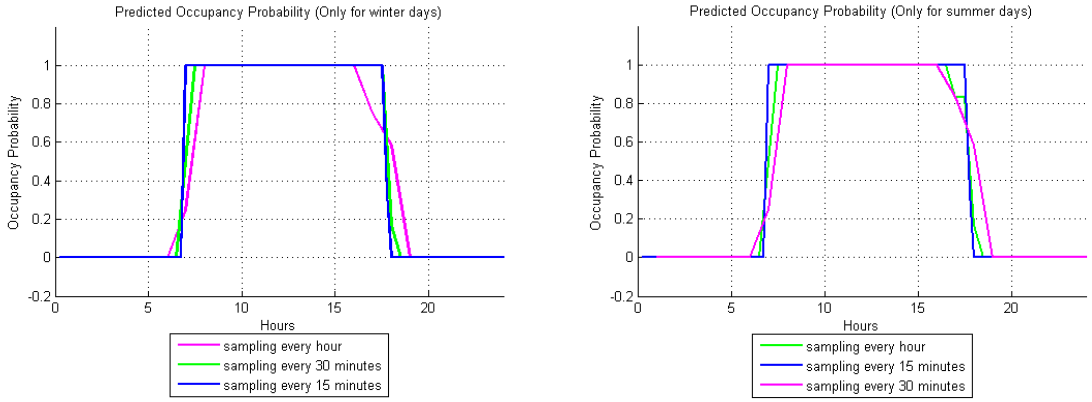
Στα πειράματα για τον Preheat αλγόριθμο χρησιμοποιήθηκαν τα ίδια δεδομένα χωρίς όμως αυτή τη φορά να γίνουν οι διαχωρισμοί των μηνών ή των ημερών. Αυτό συμβαίνει διότι η συγκεκριμένη μέθοδος για να προβλέψει το occupancy, στηρίζεται σε γνωστά δεδομένα της τρέχουσας ημέρας που έχουν καταγραφεί μέχρι την στιγμή που ξεκινά ο αλγόριθμος της πρόβλεψης. Κατά κάποιο τρόπο λοιπόν, ο αλγόριθμος εκτελεί από μόνος του τον διαχωρισμό των ημερών, επιλέγοντας να χρησιμοποιήσει τα δεδομένα των 5 ημερών που ταιριάζουν πιο πολύ με την τρέχουσα ημέρα.

Για να πραγματοποιηθούν τα πειράματα χρησιμοποιήθηκε σαν τρέχουσα ημέρα η τελευταία ημέρα του συνόλου των δεδομένων που είναι διαθέσιμα και όλες οι υπόλοιπες μέρες χρησιμοποιήθηκαν σαν το σύνολο δεδομένων από το οποίο αναζητούνται οι πλησιέστερες μέρες στην τρέχουσα ημέρα. Με μπλε κύκλους αναπαριστώνται οι προβλεπόμενες τιμές του occupancy ενώ με κόκκινα "x" αναπαριστώνται οι πραγματικές τιμές.

Πιο συγκεκριμένα, στο Σχήμα 5.5 και στο Σχήμα 5.7 παρόλο που η τιμή του threshold ισούται με 0.5 η πρόβλεψη του occupancy διαφέρει σημαντικά κυρίως μετά τις 16:00, κάτι το οποίο φαίνεται

και στις τιμές του Πίνακα 5.1. Το γεγονός αυτό οφείλεται στο ότι στο μεν Σχήμα 5.5 είναι γνωστές οι τιμές του occupancy μέχρι τις 06:00 ενώ στο δε Σχήμα 5.7 είναι γνωστό το διπλάσιο πλήθος τιμών του occupancy δηλαδή μέχρι τις 12:00. Αυτό έχει σαν αποτέλεσμα να χρησιμοποιούνται σαν πλησιέστερες μέρες διαφορετικές μέρες για την πρώτη περίπτωση και διαφορετικές μέρες για την δεύτερη περίπτωση, όπως φαίνεται και στα Σχήματα 5.6, 5.8 αντίστοιχα. Παρόμοια συμπεριφορά παρατηρείται και στις περιπτώσεις όπου επιλέγεται η τιμή του threshold να ισούται με 0.7 όπως στη σύγκριση μεταξύ των Σχημάτων 5.9 και 5.11.

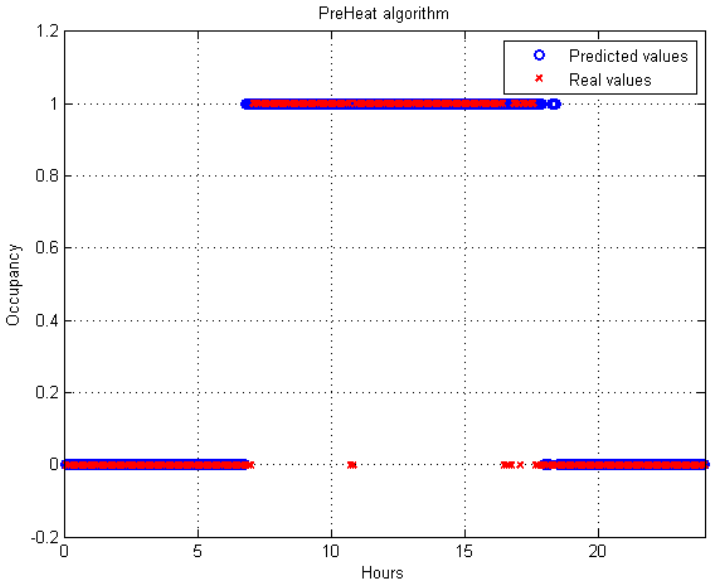
Αν συγκριθεί το Σχήμα 5.5 με το Σχήμα 5.9 όπου το πλήθος των γνωστών τιμών του occupancy είναι το ίδιο αλλά διαφέρει η τιμή του threshold και μελετηθούν οι τιμές του Πίνακα 5.1 παρατηρείται ότι η πρόβλεψη του occupancy βελτιώνεται. Αντίθετα συγκρίνοντας το Σχήμα 5.7 με το Σχήμα 5.11 παρατηρείται, με την βοήθεια του Πίνακα 5.1, ότι η πρόβλεψη του occupancy χειροτερεύει. Αυτό οφείλεται στο ότι στις δύο περιπτώσεις χρησιμοποιούνται διαφορετικές πλησιέστερες μέρες για να πραγματοποιηθεί η πρόβλεψη. Επιπλέον ένας ακόμη παράγοντας που συμβάλει σε αυτή τη διαφοροποίηση είναι ότι η κάθε μέρα διαφέρει από κάποια άλλη ιδίως μετά τις 16:00.



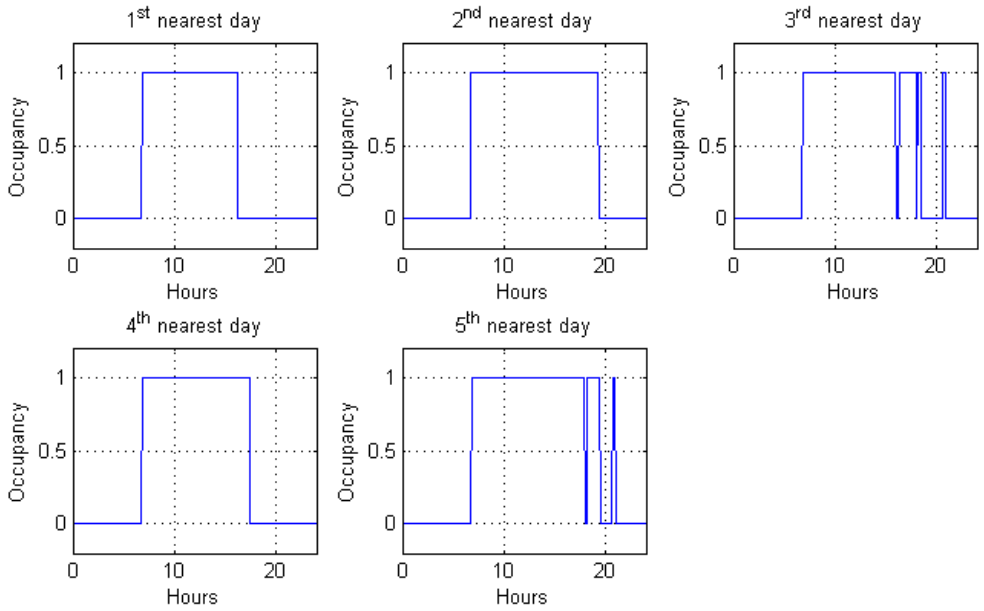
(α΄) Πιθανότητα προβλεπόμενης πληρότητας του γραφείου για μία χειμωνιάτικη ημέρα

(β΄) Πιθανότητα προβλεπόμενης πληρότητας του γραφείου για μία καλοκαιρινή ημέρα

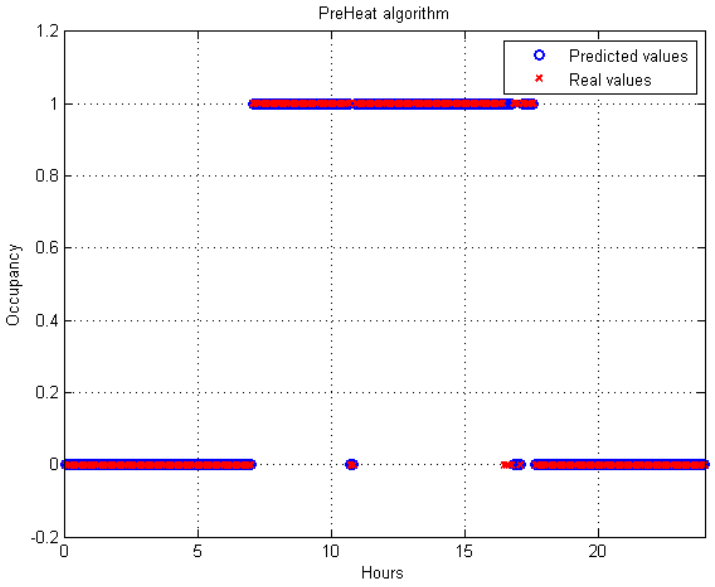
Σχήμα 5.4: Στις παραπάνω γραφικές παραστάσεις παρουσιάζεται η πιθανότητα προβλεπόμενης πληρότητας του γραφείου για δύο διαφορετικές εποχές.



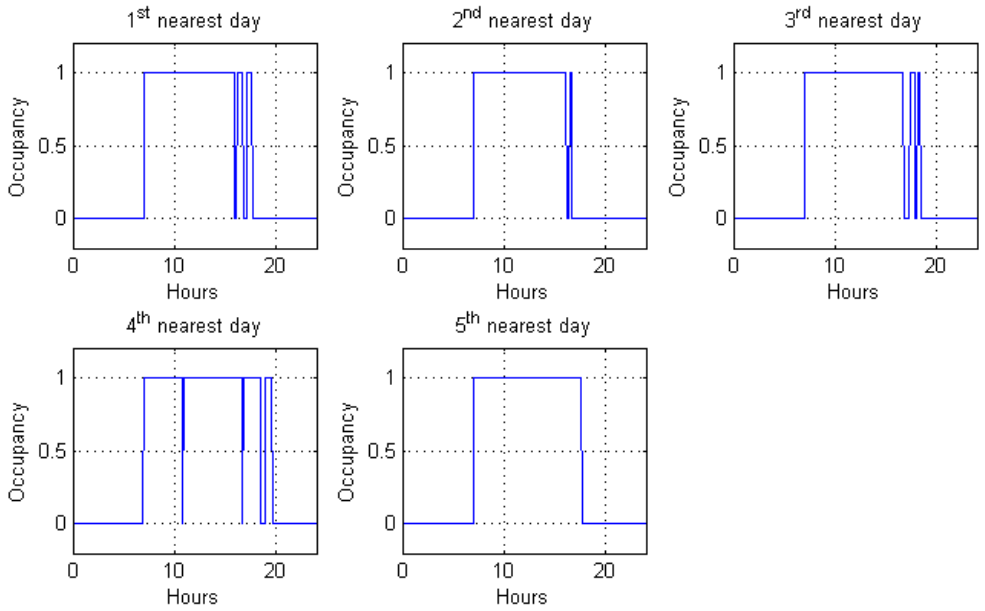
Σχήμα 5.5: Πρόβλεψη του occupancy δεδομένου ότι είναι γνωστές οι τιμές του ως τις 06:00, και threshold=0.5.



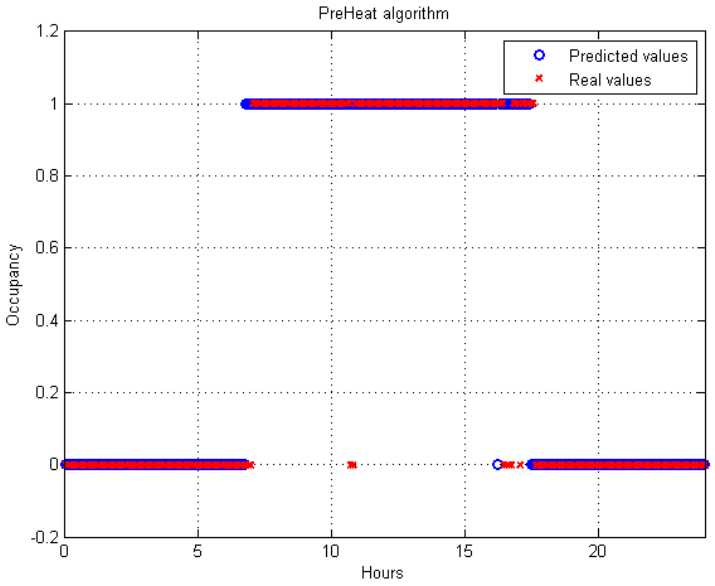
Σχήμα 5.6: Οι 5 πλησιέστερες μέρες που προέκυψαν υπολογίζοντας την απόσταση Hamming. Χρησιμοποιήθηκαν για τον υπολογισμό του occupancy στο Σχήμα 5.5.



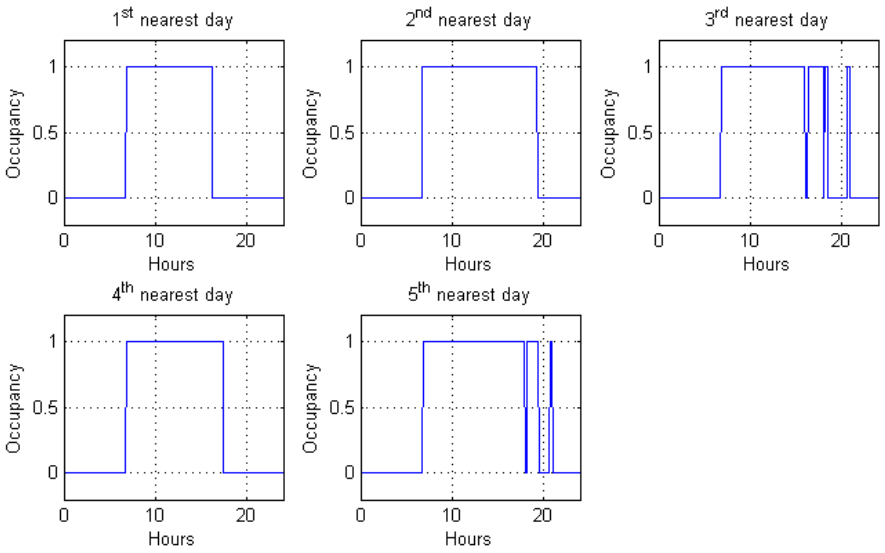
Σχήμα 5.7: Πρόβλεψη του occupancy δεδομένου ότι είναι γνωστές οι τιμές του ως τις 12:00, και threshold=0.5.



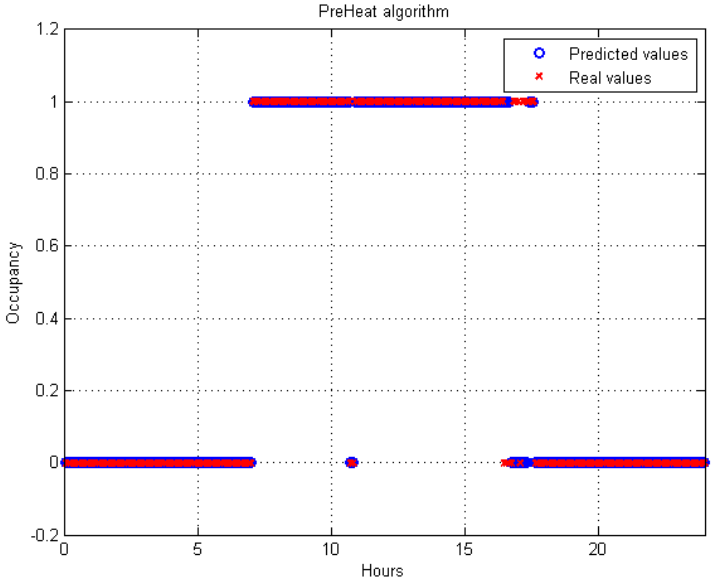
Σχήμα 5.8: Οι 5 πλησιέστερες μέρες που προέκυψαν υπολογίζοντας την απόσταση Hamming. Χρησιμοποιήθηκαν για τον υπολογισμό του occupancy στο Σχήμα 5.7.



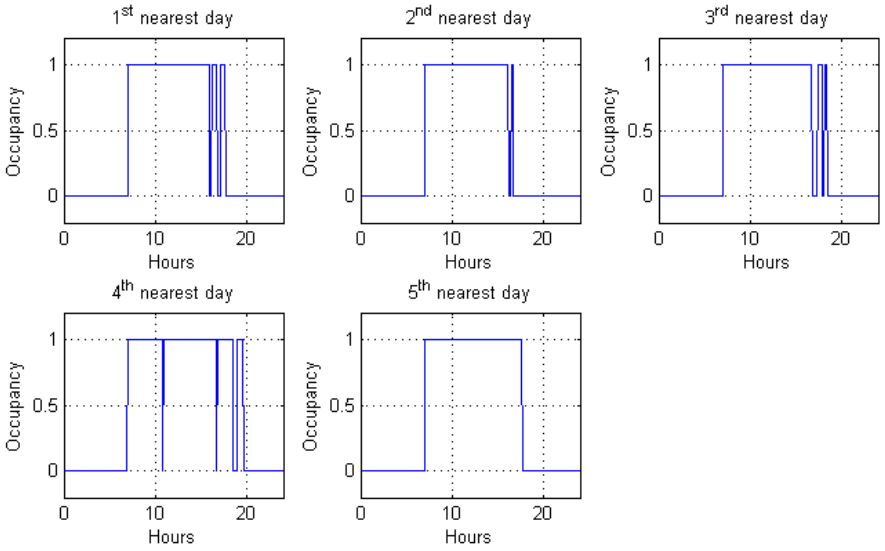
Σχήμα 5.9: Πρόβλεψη του occupancy δεδομένου ότι είναι γνωστές οι τιμές του ως τις 06:00, και threshold=0.7.



Σχήμα 5.10: Οι 5 πλησιέστερες μέρες που προέκυψαν υπολογίζοντας την απόσταση Hamming. Χρησιμοποιήθηκαν για τον υπολογισμό του occupancy στο Σχήμα 5.9.



Σχήμα 5.11: Πρόβλεψη του occupancy δεδομένου ότι είναι γνωστές οι τιμές του ως τις 12:00, και threshold=0.7.



Σχήμα 5.12: Οι 5 πλησιέστερες μέρες που προέκυψαν υπολογίζοντας την απόσταση Hamming. Χρησιμοποιήθηκαν για τον υπολογισμό του occupancy στο Σχήμα 5.11.

Πίνακας 5.1: Αποτελέσματα του συντελεστή προσδιορισμού R^2 και του μέσου τετραγωνικού σφάλματος για τα παραπάνω πειράματα.

Παραδείγματα	Όριο απόφασης			
	0.5		0.7	
Γνωστά δεδομένα ως τις 06:00	R^2	MSE	R^2	MSE
	75.71	5.9	**81.43**	**4.51**
Γνωστά δεδομένα ως τις 12:00	R^2	MSE	R^2	MSE
	90	**2.43**	87.14	3.13

Στον παραπάνω συνοπτικό πίνακα παρατίθενται τα αποτελέσματα του συντελεστή προσδιορισμού R^2 (Παράγραφος 3.8.1) καθώς και το μέσο τετραγωνικό σφάλμα (MSE) (Παράγραφος 3.8.2) των πειραμάτων. Οι τιμές που αναγράφονται είναι ποσοστού επί τοις εκατό (%).

Αυτό που παρουσιάζει ο συγκεκριμένος πίνακας είναι, αρχικά η ορθή λειτουργία των δεικτών αποτελεσματικότητας, δηλαδή όταν αυξάνεται ο R^2 μειώνεται ο MSE, καθώς και μια πιο ευδιάκριτη παρουσίαση, σε σχέση με τα σχήματα παραπάνω, της συμπεριφοράς της μεθόδου ανάλογα με την επιλογή των παραμέτρων της. Το ουσιαστικό κομμάτι που έχει να προσθέσει ο πίνακας σε όσα ειπώθηκαν στην περιγραφή των σχημάτων είναι το γεγονός ότι όταν τα γνωστά δεδομένα είναι λίγα τότε απαιτείται ορισμός μεγαλύτερου ορίου απόφασης, δηλαδή η απόφαση να γίνει πιο αυστηρή. Αυτό συμβαίνει διότι έχοντας λίγα γνωστά δεδομένα στην τρέχουσα ημέρα, η σύγκριση με τα υπάρχοντα δεδομένα επιστρέφει αποτελέσματα που τελικά να μην ταιριάζουν με την τρέχουσα ημέρα για τα υπόλοιπα δεδομένα πλην αυτών της σύγκρισης.

Κεφάλαιο 6

Συμπεράσματα

6.1 Διόρθωση Δεδομένων

Η συμπεριφορά των μεθόδων regression που χρησιμοποιήθηκαν (Polynomial regression, Local regression, Support Vector regression, Gaussian Process) σίγουρα διαφέρει λόγω της διαφορετικής προσέγγισης πού έχει η κάθε μέθοδος. Το Polynomial regression διαφέρει σημαντικά από τις υπόλοιπες μεθόδους ενώ το Local regression και Support Vector regression έχουν περισσότερες ομοιότητες καθώς στηρίζονται στο γεγονός ότι παράγουν το μοντέλο πρόβλεψης βασιζόμενες στα τοπικά δεδομένα του σημείου πρόβλεψης και όχι στο συνολικό πλήθος των δεδομένων.

Πιο αναλυτικά, το Polynomial regression δεν λειτουργεί αποδοτικά στις περισσότερες των περιπτώσεων λόγω του ότι τα δεδομένα των θερμοκρασιών δεν είναι απόλυτα γραμμικά, με αποτέλεσμα το Polynomial regression να προσφέρει μόνο μια πολύ γενική εκτίμηση των τιμών των δεδομένων, κάτι που φαίνεται στις γραφικές παραστάσεις Σχήμα 4.3, Σχήμα 4.27. Αντίθετα στο Σχήμα 4.15 παρόλο που λείπουν πολλά δεδομένα, το regression είναι πολύ αποδοτικό. Αυτό οφείλεται στο γεγονός ότι τα δεδομένα παρουσιάζουν μια γραμμική μορφή. Ένα μεγάλο μειονέκτημα της συγκεκριμένης μεθόδου είναι ότι για διαφορετικά δείγματα δεδομένων πρέπει κάθε φορά να επιλέγεται διαφορετική τάξη πολυωνύμου που να ταιριάζει περισσότερο στην καμπυλότητα των δεδομένων. Άρα λοιπόν η συγκεκριμένη μέθοδος αποδίδει καλύτερα μόνο όταν τα δεδομένα παρουσιάζουν όσο το δυνατόν γραμμικότερη καμπύλη.

Στην περίπτωση του Local regression παρατηρείται καλύτερη συμπεριφορά στο μοντέλο πρόβλεψης. Στο πρώτο πείραμα παρατηρείται ότι λόγω των έντονων αυξομειώσεων των τιμών των δεδομένων επιλέγοντας την παράμετρο εξομάλυνσης να είναι h=0.1 όπως στο Σχήμα 4.4, δηλαδή αρκετά μικρή, επιτυγχάνεται αποδοτικό regression. Όσο αυξάνεται η τιμή του h όπως στο Σχήμα 4.5 και Σχήμα 4.6, το regression ολοένα και γενικεύει. Παρόλα αυτά υπολογίζοντας το βέλτιστο h σύμφωνα με την Παράγραφο 3.3, το αποτέλεσμα του regression είναι αρκετά γενικό (Σχήμα 4.7), προσπαθώντας να αποφευχθεί τυχόν overfitting στα δεδομένα εκπαίδευσης.

Όμοια συμπεριφορά παρατηρείται και στο τρίτο πείραμα (Σχήμα 4.28, Σχήμα 4.29, Σχήμα 4.30, Σχήμα 4.31) όπου τα δεδομένα είναι περισσότερα αλλά υπάρχει ένα μεγάλο κενό δεδομένων στη μέση του δείγματος και έτσι επιλέγεται μεγαλύτερη τιμή του h ώστε να επιτευχθεί μία γενικότερη εκτίμηση που μπορεί να φανεί πιο αντιπροσωπευτική όσο αφορά τα δεδομένα που λείπουν.

Στο δεύτερο πείραμα ακολουθώντας την ίδια λογική με τα προηγούμενα, επειδή λείπουν πολλά

δεδομένα, στην ουσία η πλειοψηφία του δείγματος, επιλέγοντας μικρό h όπως στο Σχήμα 4.16 δεν επιτυγχάνεται αποδοτικό regression παρότι τα δεδομένα ακολουθούν πιο ομαλή καμπύλη σε σχέση με τα υπόλοιπα πειράματα. Η κατάσταση βελτιώνεται σημαντικά στο Σχήμα 4.17 και Σχήμα 4.18 όπου το h είναι μεγαλύτερο. Στο Σχήμα 4.19 και Σχήμα 4.20 παρατηρείται η επίδρασης του πλήθους των δεδομένων στον υπολογισμό του βέλτιστου h για την ίδια καμπυλότητα των δεδομένων.

Για την συγκεκριμένη μέθοδο regression, συμπεραίνεται ότι είναι αρκετά ευέλικτη ανεξάρτητα των χαρακτηριστικών του δείγματος των δεδομένων καθώς παρέχεται η δυνατότητα επιλογής της τιμής της παραμέτρου εξομάλυνσης ανάλογα με το τι ζητείται να επιτευχθεί και ενδείκνυται για διάφορα ειδή δεδομένων.

Το Support Vector regression είναι η πιο σταθερή μέθοδος σε σχέση με τις δύο προηγούμενες. Κύριο πλεονέκτημα της μεθόδου είναι η λεπτομερής αναζήτηση της κατάλληλης ομάδας των παραμέτρων ανάλογα με τα δεδομενα με σκοπό την βελτιστοποίηση του αποτελέσματος της συγκεκριμένης μεθόδου. Ωστόσο, όταν λείπουν μεγάλα κομμάτια δεδομένων η μέθοδος δεν αποδίδει καλά σε αυτή την περιοχή. Αυτό συμβαίνει καθώς το μοντέλο που παράγεται εξαρτάται μόνο από ένα υποσύνολο των δεδομένων εκπαίδευσης και όχι από το σύνολο τους, όπως περιγράφεται στην Παράγραφο 3.4.

Έτσι γίνεται κατανοητό πως η μέθοδος του Support Vector regression είναι κατάλληλη για διόρθωση διαφόρων ειδών δεδομένων με μόνο σημείο προσοχής όταν στα δεδομένα που επιδιώκεται διόρθωση υπάρχουν μεγάλα διαστήματα ελλειπόντων δεδομένων.

Στη μέθοδος Gaussian process παίζει σημαντικό ρόλο η χρήση της κατάλληλης covariance function ανάλογα με τα δεδομένα. Όταν λοιπόν η καμπυλότητα των δεδομένων δεν είναι ομαλή όπως στο Σχήμα 4.10 και Σχήμα 4.11 του πρώτου πειράματος, είναι προτιμότερη η χρήση της συνάρτησης Matern (Σχήμα 4.11) σε σχέση με την συνάρτηση Squared Exponential (Σχήμα 4.10). Αντίθετα όταν η καμπυλότητα των δεδομένων είναι ομαλή ανεξάρτητα από το πλήθος των δεδομένων που λείπουν, είναι προτιμότερη η χρήση της Squared Exponential συνάρτησης όπως παρουσιάζεται στο Σχήμα 4.22. Τέλος αξίζει να σημειωθεί ότι όταν το κομμάτι των δεδομένων που λείπουν είναι αρκετά μεγάλο σε σχέση με το σύνολο των δεδομένων τότε και οι δύο covariance functions παρουσιάζουν παρόμοια συμπεριφορά (Σχήμα 4.33 και Σχήμα 4.34).

Για την μέθοδο Gaussian process εν τέλη προκύπτει το συμπέρασμα πως είναι μία αξιόπιστη μέθοδος για πάσης φύσεως δεδομένα με μόνο ελάττωμα την συντηρητική προσέγγιση στα διαστήματα με πολλά ελλείποντα δεδομένα.

Όσο αφορά την μέθοδο Nearest Neighbor imputation συμπεραίνεται από όλες τις γραφικές παραστάσεις των πειραμάτων (Σχήμα 4.12-4.13, Σχήμα 4.23-4.24, Σχήμα **??**-**??**) ότι δεν είναι αξιόπιστη μέθοδος. Μπορεί να χρησιμοποιηθεί μόνο όταν το αποτέλεσμα του data correction δεν απαιτείται να είναι ακριβές ή σε δεδομένα οι τιμές των οποίων να παραμένουν σχεδόν ίδιες κατά το πέρασμα του χρόνου.

Η τελευταία μέθοδος που χρησιμοποιήθηκε στην παρούσα διπλωματική είναι αυτή του Neural Network fitting. Στη συγκεκριμένη μέθοδο όταν τα δεδομένα έχουν ομαλή καμπυλότητα όπως στο Πείραμα 2 ή σχετικά ομαλή καμπυλότητα όπως στο Πείραμα 3, τότε συνιστάται ο αλγόριθμος Bayesian regularization (Σχήμα 4.26 και Σχήμα **??**) και όχι ο αλγόριθμος Levenberg-Marquardt (Σχήμα 4.25 και Σχήμα **??**), χωρίς να είναι απαγορευτική η χρήση του αλλά θα χρειαστεί ρύθμιση η παράμετρος d που τον χαρακτηρίζει. Αντίθετα όταν παρουσιάζονται έντονες αυξομειώσεις στα δεδομένα είναι αποτελεσματικότερος ο αλγόριθμος Levenberg-Marquardt χάρη στη δυνατότητα ρύθμισης της

d που ορίζει πόσες φορές παραγωγίζεται το μοντέλο με σκοπό να προσαρμοστεί σε μη ομαλά δεδομένα. Αρνητικό αυτής της παραμέτρου παραμένει το γεγονός της μειωμένης δυνατότητας σύγκρισης μεταξύ διαφορετικών εκτιμήσεων, ίδιων δεδομένων, λόγω στην τυχαιότητα με την οποία επιλέγονται τα αρχικά βάρη στην εκπαίδευση των νευρώνων.

6.2 Πρόβλεψη Πληρότητας

6.2.1 Αλγόριθμος SmartThermostat

Αρχικά παρατηρώντας το Σχήμα 5.1 γίνεται αντιληπτό ότι οι εργαζόμενοι στο γραφείο πηγαίνουν από τις 06:00 ως τις 07:00 και μετά τις 15:30 λείπουν όλοι για σύντομα χρονικά διαστήματα και επανέρχεται κάποιος ξανά για μικρό χρονικό διάστημα ως περίπου στις 21:00 που φεύγουν και δεν επιστρέφουν ξανά μέχρι το επόμενο πρωί, ανεξάρτητα από ποία μέρα του χρόνου είναι.

Στη συνέχεια συγκρίνοντας τα σχήματα μεταξύ τους (5.1, 5.2, 5.3, 5.4) παρατηρείται ότι ο αλγόριθμος SmartThermostat κάνει πολύ καλή πρόβλεψη του occupancy παρόλο που δεν καταφέρνει να προβλέψει τις συχνές μεταβολές που παρατηρούνται μετά τις 15:30. Ένα ακόμη συμπέρασμα που προκύπτει από την παρατήρηση-σύγκριση των παραπάνω σχημάτων είναι ότι όσο πιο μικρό είναι το διάστημα δειγματοληψίας τόσο πιο απόλυτη-ακριβής γίνεται η πρόβλεψη ενώ όσο μεγαλώνει το διάστημα δειγματοληψίας τότε η πρόβλεψη γίνεται πιο γενική καταφέρνοντας να παρουσιάσει έστω και λίγο την διαφοροποίηση του occupancy μετά τις 15:00.

Από τις γραφικές παραστάσεις Σχήμα 5.2 και Σχήμα 5.3, γίνεται αντιληπτό ότι για τα δεδομένα που χρησιμοποιήθηκαν ώστε να δημιουργηθούν τα παραπάνω πειράματα, ο διαχωρισμός τους σε ίδιες μέρες δεν παρουσιάζει κάποια έντονη διαφοροποίηση στην πρόβλεψη. Αυτό προφανώς συμβαίνει γιατί στο συγκεκριμένο γραφείο από το οποίο λήφθηκαν τα δεδομένα έχουν σταθερό ωράριο, δηλαδή πηγαίνουν στο γραφείο περίπου στις 07:00 και αποχωρούν μετά τις 16:00 όπως φαίνεται και στα σχήματα. Παρόμοια συμπεριφορά παρατηρείται και στις γραφικές παραστάσεις Σχήμα 5.4 όπου πάλι προκύπτει το συμπέρασμα ότι ο διαχωρισμός μεταξύ χειμερινών και καλοκαιρινών ημερών δεν ήταν απαραίτητος για τα συγκεκριμένα δεδομένα.

6.2.2 Αλγόριθμος PreHeat

Παρατηρώντας τις γραφικές παραστάσεις στην Παράγραφο 5.3 είναι φανερό ότι η πρόβλεψη της πληρότητας επηρεάζεται σημαντικά από τις τιμές που θα επιλεγούν για δύο κύριες παραμέτρους. Η μία παράμετρος είναι το όριο (threshold) της πιθανότητας πρόβλεψης που αν ξεπεραστεί, ο αλγόριθμος προβλέπει ότι ο χώρος που μελετάται είναι occupied. Η δεύτερη παράμετρος είναι μέχρι ποια χρονική στιγμή της ημέρας έχουν παρατηρηθεί τιμές του occupancy.

Ολοκληρώνοντας προκύπτει το συμπέρασμα πως ο Preheat αλγόριθμος επηρεάζεται σημαντικά από την τυχαιότητα που παρουσιάζει η κάθε ημέρα, όσο αφορά την τιμή του occupancy, κυρίως μετά τις 16:00, λόγω του ότι πολύ σπάνια μοιάζουν τα αποτελέσματα του occupancy για διαφορετικές μέρες μετά τις 16:00.

6.2.3 Γενίκευση Συμπερασμάτων

Στην παρούσα διπλωματική εργασία τα δεδομένα που χρησιμοποιήθηκαν, αρχικά για τις μεθόδους διόρθωσεις δεδομένων ήταν η θερμοκρασία του χώρου και έπειτα για τις μεθόδους πρόβλεψης πληρότητας ήταν τα δεδομένα παρουσίας ατόμων στο κτήριο (πληρότητα-occupancy). Χρησιμοποιήθηκαν αυτά τα δεδομένα καθώς είναι τα πλέον κατάλληλα δεδομένα για την δοκιμή των μεθόδων που μελετήθηκαν. Επιπλέον η θερμοκρασία ενός χώρου είναι από τα πιο ευμετάβλητα μεγέθη ενός κτηρίου καθώς επηρεάζεται από πολλούς παράγοντες, γι' αυτό και επιλέχθηκε στα πειράματα καθώς μπορεί να προσφέρει πολύ διαφορετικά δείγματα δεδομένων.

Το γεγονός ότι στα πειράματα χρησιμοποιήθηκαν μόνο τα δεδομένα της θερμοκρασίας και της πληρότητας δεν σημαίνει πως οι μέθοδοι που μελετήθηκαν λειτουργούν ή αποδίδουν μόνο για αυτά τα δύο μεγέθη. Προφανώς λειτουργούν και για όλα τα άλλα μεγέθη που χαρακτηρίζουν τις συνθήκες ενός κτηρίου όπως η υγρασία, η φωτεινότητα, η ποσότητα διοξειδίου του άνθρακα (CO_2) κλπ, καθώς τα υπόλοιπα μεγέθη πέραν της θερμοκρασίας είναι πιο "απλά" με την έννοια ότι η μεταβολή τους κατά την διάρκεια της ημέρας είναι πιο ομαλή. Όμοια και οι μέθοδοι πρόβλεψης πληρότητας θα μπορούσαν κάλλιστα να χρησιμοποιηθούν, ίσως με μικρές αλλαγές, για την πρόβλεψη μετεωρολογικών παραμέτρων οι οποίες σχετίζονται με τις απαιτήσεις σε θέρμανση ή ψύξη ενός κτηρίου.

Τα δεδομένα που χρησιμοποιήθηκαν όπως έχει ήδη ειπωθεί αφορούν πραγματικά δεδομένα του κτηρίου του τεχνολογικού κέντρου Cartif. Στην θέση του συγκεκριμένου κτηρίου θα μπορούσε να βρίσκεται οποιοδήποτε άλλο κτήριο χρησιμοποιεί ένα τρόπο διατήρησης αντίστοιχων δεδομένων. Η μόνη διαφοροποίηση θα ήταν στις τιμές των δεδομένων καθώς σχετίζονται με το γεωγραφικό σημείο που βρίσκεται κάθε κτήριο χωρίς όμως το γεγονός αυτό να επηρεάζει τη λειτουργία και την απόδοση των μεθόδων.

6.2.4 Μελλοντικές Προεκτάσεις

Μελλοντικά θα μπορούσε να αναπτυχθεί μία εφαρμογή, η οποία χρησιμοποιώντας την παρούσα μελέτη των μεθόδων διόρθωσης δεδομένων, να είναι σε θέση να επιλέγει την κατάλληλη μέθοδο βάσει χαρακτηριστικών των δεδομένων όπως η ομαλότητα τους, το μέγεθος τους, η πυκνότητα τους και το πλήθος των δεδομένων που λείπουν. Σκοπός της εφαρμογής θα είναι η γρήγορη και άμεση συμβολή της στην επεξεργασία και αξιολόγηση των δεδομένων ενός κτηρίου.

Από την μεριά της πρόβλεψης πληρότητας ενός κτηρίου, μελλοντική προέκταση μπορεί να θεωρηθεί η πρόβλεψη της συμπεριφοράς των ενοίκων που επηρεάζει άμεσα και έμμεσα την ενεργειακή συμπεριφορά του κτηρίου μέσα από το άνοιγμα-κλείσιμο των παραθύρων, ενεργοποίηση-απενεργοποίηση ή ρύθμιση των φώτων, την ενεργοποίηση-απενεργοποίηση του εξοπλισμού γραφείου, ενεργοποίηση-απενεργοποίηση θέρμανσης, αερισμού και κλιματισμού.

Βιβλιογραφία

[1] Hidden markov models (hmm). http://www.mathworks.com/help/stats/hidden-markov-models-hmm.html#bq_i1wh.

[2] Impute missing data using nearest-neighbor method. http://www.mathworks.com/help/bioinfo/ref/knnimpute.html.

[3] Local linear kernel regression. http://www.mathworks.com/matlabcentral/fileexchange/19564-local-linear-kernel-regression.

[4] Y. Agarwal, B. Balaji, S. Dutta, R. Gupta, and T.Weng. Duty-cycling buildings aggressively: The next frontier in hvac control. *Proceedings of the 10th International Conference on Information Processing in Sensor Networks (IPSN), Pages 246 - 257, Chicago, IL*, 2011.

[5] Paul D. Allison. Modern methods for missing data. https://www.amstat.org/sections/srms/webinarfiles/ModernMethodWebinarMay2012.pdf.

[6] Paul D. Allison. Multiple imputation for missing data: A cautionary tale. Technical report, Sociology Department, University of Pennsylvania.

[7] Paul D. Allison. Missing data. In Roger E. Millsap and CA: Sage Publications Inc. Alberto Maydeu-Olivares. Thousand Oaks, editors, *The SAGE Handbook of Quantitative Methods in Psychology*, pages 72–83. 2009.

[8] Paul D. Allison. Handling missing data by maximum likelihood. Technical report, 2012.

[9] Mauricio A. Alvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: a review. *Foundations and Trends® in Machine Learning Journal, Volume 4 Issue 3, Pages 195-266.*

[10] Ibrahim Berkan Aydilek and Ahmet Arslan. A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks. *International Journal of Innovative Computing, Information and Control*, 8(7(A)), 2012.

[11] BaaS. Building as a service, 2012. http://www.baas-project.eu/.

[12] Gustavo E. A. P. A. Batista and Maria Carolina Monard. A study of k-nearest neighbour as an imputation method. Technical report, Institute of Mathematics and Computer Science, University of Sao Paulo, 2002.

[13] Dimitri P. Bertsekas. *Constrained Optimazation and Lagrance Multiplier Methods*. Athena Scientific Belmont, MA, 2008. ISBN: 1-886529-04-3.

[14] D.P. Bertsekas, WW Hager, and OL Mangasarian. *Nonlinear programming*. Athena Scientific Belmont, MA, 1999. ISBN:1886529000.

[15] Christopher M. Bishop. Neural networks for pattern recognition. Technical report, Department of Computer Science and Applied Mathematics, Aston University, Birmingham, UK, 1995.

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006. ISBN:0387310738.

[17] Andrian W. Bowman and Adelchi Azzalini. *Applied Smoothing Techniques for Data Analysis*. Oxford Science Publications, 1997. ISBN: 0198523963.

[18] Phillip Boyle. *Gaussian Processes for Regression and Optimisation*. PhD thesis, Victoria University of Wellington, 2007.

[19] R. Dodier, G. Henze, D. Tiller, and X. Guo. Building occupancy detection through sensor belief networks. *Energy and Buildings*, Volume 38, Issue 9:1033–1140, 2006.

[20] Justin Domke and Linwei Wang. Overfitting and model selection. `http://phd.gccis.rit.edu/discovery/proj1/Overfitting.pdf`.

[21] Norman R. Draper and Harry Smith. *Applied Regression Analysis*. Wiley, 1998. ISBN: 978-0-471-17082-2.

[22] Echelon. *i.LON® 100 e3 User's guide*, 2006.

[23] Echelon. The lonworks® protocol. 2013. `http://www.echelon.com/technology/lonworks/lonworks-protocol.htm`.

[24] V. Erickson, M. Carreira-Perpiñán, and A. Cerpa. Occupancy based system for efficient reduction of hvac energy. *Proceedings of the 10th International Conference on Information Processing in Sensor Networks (IPSN), Pages 258 - 269, Chicago, IL*, 2011.

[25] Aly Farag and Refaat M Mohamed. Regression using support vector machines: Basic foundations. Technical report, Computer Vision and Image Processing Laboratory, Electrical and Computer Engineering Department, University of Louisville, 2004.

[26] Miguel Á. García, Cristina de Torre, Andrés Macía, José L. Hernández, César Valmaseda, Javier Martín, Juan Rodríguez, Dimitrios Rovas, Giorgos Kontes, Giorgos Giannakis, and Kyriakos Katsigarakis. Deliverable 6.1: Identification and definition of baas demonstration buildings. *BaaS project*, 2013.

[27] Miguel Á. García, Cristina de Torre, Andrés Macía, José L. Hernández, César Valmaseda, Javier Martín, Óscar Hidalgo, Kyriakos Katsigarakis, Giorgos Giannakis, Dimitrios Rovas, and Juan Rodríguez. Deliverable 6.2: Operative pilots after adapting. *BaaS project*, 2013.

[28] M. Gupta, S. S. Intille, and K. Larson. Adding gps-control to traditional thermostats: An exploration of potential energy savings and design challenges. *Proceedings of the 7th International Conference on Pervasive Computing Pages 95 - 114, Nara, Japan*, 2009.

[29] Honeywell. Symmetre® r310. `http://www.zarifopoulos.com/files/SymmetrE%20R310.pdf`.

[30] J. Krumm and A. J. Brush. Learning time-based presence probabilities. *Proceedings of the 9th International Conference, Pervasive, Pages 79 - 96, San Francisco, USA*, 2011.

[31] C.J. Lin, C. Chang, and C. Hsu. A practical guide to support vector classification. *National Taiwan University*, 2004. `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

[32] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data, 2nd Edition*. Wiley, 2002. ISBN: 978-0-471-18386-0.

[33] Jiakang Lu, Tamim Sookoor, Vijay Srinivasan, Ge Gao, Brian Holben, John Stankovic, Eric Field, and Kamin Whitehouse. The smart thermostat: Using occupancy sensors to save energy in homes. *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems Pages 211-224, Zurich, Switzerland*, 2010.

[34] Microsoft. Sql server. `http://www.microsoft.com/en-us/server-cloud/products/sql-server/`.

[35] Fulufhelo Vincent Nelwamondo. *Computational Intelligence Techniques for Missing Data Imputation*. PhD thesis, University of the Witwatersrand, Johannesburg, 2006.

[36] Liqiang Pan and Jianzhong Li. K-nearest neighbor based missing data estimation algorithm in wireless sensor networks. *Wireless Sensor Network journal, Vol.2 No.2, Pages 115-122*, 2010.

[37] Yongsong Qin, Shichao Zhang, Xiaofeng Zhu, Jilian Zhang, and Chengqi Zhang. Pop algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases. *Expert Systems with Applications: An International Journal, Volume 36 Issue 2, Pages 2794-2804*, 2009.

[38] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE (Volume:77 , Issue: 2), Pages 257-286*, 1989.

[39] Ananth Ranganathan. The levenberg-marquardt algorithm. Technical report, 2004.

[40] Carl Edward Rasmussen and Hannes Nickisch. *The GPML Toolbox version 3.4*, 2014.

[41] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005. ISBN: 026218253X.

[42] Carl Edward Rasmussen and Christopher K. I. Williams. Examples of covariance functions. In *Gaussian Processes for Machine Learning*, chapter 4.2. 2006.

[43] Carl Edward Rasmussen and Christopher K. I. Williams. Function-space view. In *Gaussian Processes for Machine Learning*, chapter 2.2. 2006.

[44] John O. Rawlings, Sastry G. Pantula, and David A. Dickey. *Applied Regression Analysis: A Research Tool, Second Edition*. Springer, 1998. ISBN: 0-387-98454-2.

[45] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell. Nextplace: A spatio-temporal prediction framework for pervasive systems. *Proceedings of the 9th international conference on Pervasive computing, Pages 152-169 , San Francisco, USA*, 2011.

[46] Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. Technical report, Pennsylvania State University, 2002.

[47] Bernhard Scholkopf and Alexander J. Smola. *Learning with kernels: Support Vector Machine, Regularization, Optimization and Beyond*. The MIT Press, 2002. ISBN: 9780262194754.

[48] James Scott, A.J. Bernheim Brush, John Krumm, Brian Meyers, Mike Hazas, Steve Hodges, and Nicolas Villar. Preheat: Controlling home heating using occupancy prediction. *Proceedings of the 13th international conference on Ubiquitous computing Pages 281-290, Beijing, China*, 2011.

[49] Alex J. Smola and Bernhard Scholkopf. A tutorial on support vector regression. Technical report, NeuroCOLT Technical Report Series, 1998.

[50] S. Tominaga, M. Shimosaka, R. Fukui, and T. Sato. A unified framework for modeling and predicting going-out behavior. *Proceedings of the 10th international conference on Pervasive Computing Pages 73-90 , Newcastle, UK*, 2012.

[51] Stef van Buuren. *Flexible Imputation of Missing Data*. CRC Press, Taylor and Francis Group, 2012. ISBN: 9781439868249.

[52] Fulufhelo V.Nelwamondo, Dan Golding, and Tshilidzi Marwala. A dynamic programming approach to missing data estimation using neural networks. *Information Sciences:an International Journal, Volume 237, Pages 49-58*, 2009.

[53] Dennis D. Wackerly, William Mendenhall, and Richard L. Scheaffer. *Mathematical Statistics with Applications*. Belmont, CA, USA: Thomson Higher Education, 2008. ISBN: 978-0-495-38508-0.

[54] Y. Ye, Y. Zheng, Y. Chen, J. Feng, and X. Xie. Mining individual life pattern based on location history. *Proceedings of the 10th International Conference on Mobile Data Management: Systems, Services and Middleware, Pages 1-10, Taipei, Taiwan*, 2009.