# ΣΥΣΧΕΤΙΣΗ ΔΕΔΟΜΕΝΩΝ ΣΕ ΣΥΝΘΗΚΕΣ ΑΒΕΒΑΙΟΤΗΤΑΣ

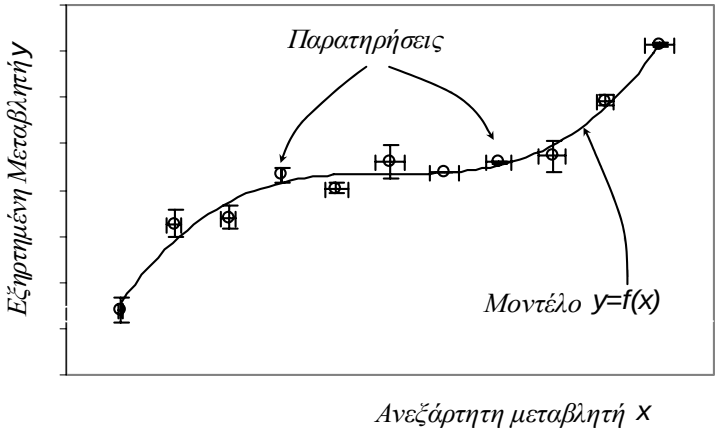
**Ε. Μαθιουλάκης, Β. Μπελεσιώτης**
*Εργαστήριο Ηλιακών & άλλων Ενεργειακών Συστημάτων - ΕΚΕΦΕ «Δημόκριτος»*
*15310 Αγ. Παρασκευή Αττικής*

**ΠΕΡΙΛΗΨΗ**
Στην εργασία αυτή παρουσιάζεται μια γενική μεθοδολογική προσέγγιση για την αντιμετώπιση του γενικού προβλήματος της γραμμικής συσχέτισης δεδομένων μεταξύ δύο ή και περισσοτέρων μεταβλητών, όταν είναι διαθέσιμες παρατηρήσεις των μεταβλητών αυτών σε συνθήκες αβεβαιότητας. Το ζητούμενο είναι ο ποσοτικός και ποιοτικός προσδιορισμός των συντελεστών ενός προεπιλεγμένου μοντέλου, στη βάση όχι μόνο των διαθέσιμων τιμών των ανεξάρτητων μεταβλητών αλλά και των αβεβαιοτήτων που χαρακτηρίζουν τις τιμές αυτές. Η προτεινόμενη προσέγγιση, βασισμένη στη χρήση των σταθμισμένων ελαχίστων τετραγώνων, εκτός από μια πιο ρεαλιστική αποτίμηση των συντελεστών του μοντέλου, οδηγεί στην εκτίμηση των αβεβαιοτήτων στους συντελεστές αυτούς και, κατά συνέπεια, στην εκτίμηση των αβεβαιοτήτων στην εξαρτημένη μεταβλητή. Επιπλέον, δίνει τη δυνατότητα να ελεγχθεί η καταλληλότητα του ίδιου του μοντέλου, συγκρίνοντας τις αποκλίσεις του από τις εμπειρικές παρατηρήσεις με τις πειραματικές αβεβαιότητες.

## 1. Εισαγωγή

Το πρόβλημα που εξετάζεται στην εργασία αυτή αφορά τη συσχέτιση μεταξύ δύο μεταβλητών ή δύο ομάδων μεταβλητών, στην προοπτική της επιβεβαίωσης ενός μοντέλου συσχέτισης το οποίο θα μπορεί να χρησιμοποιηθεί στη συνέχεια για να προβλεφθούν οι τιμές της μιας ομάδας όταν είναι γνωστές οι τιμές της άλλης. Οι τιμές των μεταβλητών είναι είτε αποτελέσματα μετρήσεων, είτε μετασχηματισμοί δεδομένων μετρήσεων.
Η συσχέτιση δεδομένων συνιστά ένα από τα σημαντικότερα προβλήματα κατά την επεξεργασία των εμπειρικών παρατηρήσεων και, ειδικότερα, στη μετρολογία [1]. Ως χαρακτηριστικό παράδειγμα αναφέρεται η διακρίβωση, όπου το ζητούμενο είναι η συσχέτιση των υψηλότερης μετρολογικής ποιότητας τιμών αναφοράς με τις αντίστοιχες ενδείξεις του υπό διακρίβωση οργάνου. Γενικεύοντας εξάλλου την έννοια της διακρίβωσης στη εύρεση της αλληλεξάρτησης μεταξύ μεταβλητών οι οποίες συνδέονται - ή υποθέτουμε ότι μπορεί να συνδέονται - μέσω μιας σχέσης αιτιότητας, η συσχέτιση δεδομένων αποτελεί μια ευρέως διαδεδομένη δραστηριότητα τόσο στις φυσικές όσο και στις κοινωνικές ή οικονομικές επιστήμες [2]. Παραμένοντας στον τομέα της μετρολογίας, μια άλλη δραστηριότητα που εμπλέκει συσχέτιση εμπειρικών δεδομένων συναντάται στις δοκιμές χαρακτηρισμού προϊόντων, όπου συνήθως είναι διαθέσιμο ένα γενικό μοντέλο συμπεριφοράς και αναζητούνται οι συντελεστές του μοντέλου που αρμόζουν στο κάθε ειδικό προϊόν.
Η συνήθης προσέγγιση είναι η γνωστή σε όλους μέθοδος των ελαχίστων τετραγώνων η οποία μπορεί, υπό ορισμένες προϋποθέσεις, να είναι αποτελεσματική για το σκοπό που εφαρμόζεται. Αυτό που ενδιαφέρει εδώ είναι η αποτελεσματικότητα της συσχέτισης, τόσο όσον αφορά τη δυνατότητά της να επεξηγήσει ή να επιβεβαιώσει τη σχέση αιτιότητας που αναμένεται ή εικάζεται ότι συνδέει τα δεδομένα, όσο και τη δυνατότητα χρησιμοποίησης της - επιβεβαιωμένης πια - συσχέτισης για την πρόβλεψη της μελλοντικής συμπεριφοράς

της μιας ομάδας μεταβλητών στη βάση της συμπεριφοράς της άλλης ομάδας. Η ποιότητα αυτή εξαρτάται προφανώς από την ποιότητα του μοντέλου συσχέτισης που υιοθετείται, συμπεριλαμβανομένης της κατεύθυνσης της αιτιότητας που συνδέει της μεταβλητές μεταξύ τους. Εξαρτάται όμως και από το πόσο σίγουροι είμαστε για τα δεδομένα που χρησιμοποιούνται για την εύρεση της συσχέτισης, δηλαδή από την αβεβαιότητα που τα χαρακτηρίζει (σχήμα 1), γεγονός που συχνά αγνοείται στην πράξη, ειδικότερα όταν οι αβεβαιότητες αυτές αφορούν τόσο τις εξαρτημένες όσο και τις ανεξάρτητες μεταβλητές.




*Σχήμα 1: Συσχέτιση δεδομένων σε συνθήκες αβεβαιότητας*

Στην παρούσα εργασία εξετάζεται κυρίως αυτή η παράμετρος - δηλαδή η συσχέτιση δεδομένων σε συνθήκες αβεβαιότητας - και προτείνονται τεχνικές για την βέλτιστη και ρεαλιστική υλοποίησή της. Η σημασία του προβλήματος αυτού γίνεται εμφανής εάν αναλογιστούμε ότι, σύμφωνα με τις τρέχουσες αντιλήψεις περί ποιότητας της μέτρησης, η αβεβαιότητα συνιστά καθολικό χαρακτηριστικό των πειραματικών αποτελεσμάτων. Θα ήταν προφανώς ανακόλουθο να αγνοηθεί η ύπαρξη αβεβαιότητας στα συσχετιζόμενα δεδομένα, όταν το ζητούμενο δεν είναι μόνο η εύρεση της σχέσης που τα συνδέει, αλλά και να υπάρχει μια εκτίμηση για την ποιότητα της σχέσης αυτής και την ποιότητα των αποτελεσμάτων που θα παράγονται από τη μελλοντική της χρήση.

Η συνήθης δραστηριότητα κατά την επεξεργασία δεδομένων που προέρχονται από παρατηρήσεις, αφορά την εύρεση μιας σχέσης υπό μορφή συνάρτησης $y=f(x_1, x_2,...x_K)$ μεταξύ μιας *εξαρτημένης* μεταβλητής $y$ και μιας ή περισσοτέρων *ανεξάρτητων* μεταβλητών $x_1, x_2,...x_K$. Τα δεδομένα της κάθε παρατήρησης συνίστανται εν γένει σε τιμές τόσο της εξαρτημένης, όσο και των ανεξάρτητων μεταβλητών. Το σύνολο των $N$ παρατηρήσεων οδηγεί επομένως σε ένα σύνολο τιμών $y_i, x_{1i}, x_{2i},...x_{Ki}$, για $i=1,...N$

Ανάλογα με το βαθμό γνώσης του μοντέλου συσχέτισης διακρίνονται οι ακόλουθες περιπτώσεις:

- Όταν δεν υπάρχει γνωστό *a priori* μοντέλο και γίνεται προσπάθεια να «εκμαιευτεί» από τα ίδια τα δεδομένα μια σχέση *("let the data to speak for themselves")*. Στην περίπτωση αυτή, η συζήτηση της οποίας ξεφεύγει από το πλαίσιο της παρούσας εργασίας, η διαδικασία συσχέτισης θέτει πολλά ερωτήματα τα οποία δεν είναι πάντα εύκολο να απαντηθούν.
- Συχνά, και οπωσδήποτε πιο συχνά στη μετρολογία, η μορφή της συνάρτησης $y=f(x_1, x_2,...x_K)$ θεωρείται εκ των προτέρων γνωστή και επιλέγεται ως το κατάλληλο θεωρητικό μοντέλο που αντιστοιχεί στη αντίστοιχη φυσική διεργασία, για παράδειγμα ο νόμος του Ohm όταν συσχετίζονται τάση στα άκρα μιας αντίστασης και ένταση

ρεύματος που τη διαπερνά ή ακόμα η πυκνότητα ως συνάρτηση του όγκου και της μάζας ενός υλικού. Στην περίπτωση αυτή το ζητούμενο είναι ο προσδιορισμός των συντελεστών της εξίσωσης, οι οποίοι συντελεστές χαρακτηρίζουν το κάθε συγκεκριμένο προϊόν ή διεργασία.

– Σε μερικές περιπτώσεις η εξίσωση συσχέτισης μπορεί να επιλεγεί ανάμεσα σε ένα περιορισμένο αριθμό υποψήφιων συναρτήσεων οι οποίες μπορούν όλες να θεωρηθούν κατάλληλες, η κάθε μια με διαφορετικό βαθμό αξιοπιστίας ή πολυπλοκότητας. Το ζητούμενο εδώ είναι, πέρα από τον προσδιορισμό των συντελεστών της κάθε μιας από αυτές, η αξιολόγηση των επιδόσεων τους, καθώς και της καταλληλότητάς τους για το συγκεκριμένο φαινόμενο (*goodness of fit*). Και στην περίπτωση αυτή, όπως και στην προηγούμενη (γνωστό ή γνωστά μοντέλα), οι αποκλίσεις των δεδομένων από το μοντέλο μπορούν να αποδοθούν σε σφάλματα των μετρήσεων και σε ατέλειες του ίδιου του μοντέλου.

Στη συνέχεια εξετάζονται μοντέλα συσχέτισης τα οποία μπορούν να διατυπωθούν ως γραμμική συνάρτηση *M* άγνωστων συντελεστών, της μορφής:

$$y = f(x_1, x_2, ..., x_M) = \sum_{j=1}^{M} a_j f_j \qquad (1)$$

όπου όμως οι επιμέρους συναρτήσεις $f_j = f_j(x_1, x_2, ..., x_K)$ μπορούν να είναι μη γραμμικές ως προς μία ή περισσότερες από τις μεταβλητές $x_1, x_2, ..., x_K$. Να σημειωθεί ότι ο αριθμός *K* των μεταβλητών δεν είναι αναγκαστικά ίσος με τον αριθμός *M* των συντελεστών.

Το τελικό ζητούμενο είναι:

- Ο προσδιορισμός των *M* άγνωστων συντελεστών $a_j$, όταν ο αριθμός *N* των παρατηρήσεων είναι μεγαλύτερος από τον αριθμό *M* των αγνώστων συντελεστών (η περίπτωση *N=M* αφορά επίλυση ενός συστήματος Ν εξισώσεων και Ν αγνώστους το οποίο επιλύεται κατά τα γνωστά, ενώ αυτή που *N<M* είναι διαφορετική και θέτει άλλου είδους προβλήματα τα οποία ξεφεύγουν από τα πλαίσια της παρούσας εργασίας).
- Η εκτίμηση της αβεβαιότητας που χαρακτηρίζει τους συντελεστές αυτούς αλλά και τα αποτελέσματα στα οποία θα οδηγήσει η μελλοντική χρήση της συνάρτησης $y = f(x_1, x_2, ..., x_K)$.
- Η καταλληλότητα του μοντέλου για να εξηγήσει τα δεδομένα των εμπειρικών παρατηρήσεων.

## 2. Η κλασική μέθοδος των ελαχίστων τετραγώνων

Η κλασική μέθοδος των ελαχίστων τετραγώνων (*least square LS*), όπως διατυπώθηκε αρχικά από τον Gauss, βασίζεται στην εύρεση ενός σετ συντελεστών $a_1, a_2, ...a_M$ το οποίο ελαχιστοποιεί το άθροισμα των τετραγώνων των διαφορών $e_i = y_i - f(x_{1i}, x_{2i}, ..., x_{Ki})$ ανάμεσα στην τιμή $y_i$ που προέκυψε κατά την παρατήρηση και αυτή που υπολογίζεται από το μοντέλο συσχέτισης:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} [y_i - f(x_{1i}, x_{2i}, ..., x_{Ki})]^2 \qquad (2)$$

Να σημειωθεί ότι η επιλογή του κριτηρίου αυτού είναι αυθαίρετη, δεδομένου ότι θα μπορούσε να έχει επιλεγεί ένα άλλο κριτήριο, για παράδειγμα η ελαχιστοποίηση του αθροίσματος των απολύτων τιμών των αποκλίσεων $|e_i| = |y_i - f(x_{1i}, x_{2i}, ..., x_{Ki})|$.

Πρέπει να τονιστεί ότι η προσέγγιση αυτή βασίζεται στην υπόθεση ότι οι αποκλίσεις του μοντέλου από τις παρατηρήσεις δεν οφείλονται σε ενδεχόμενα πειραματικά σφάλματα στις τιμές των $x_{ki}$ και $y_i$, αλλά μόνο στην αδυναμία του μοντέλου να εκφράσει με ακρίβεια τις πειραματικές τιμές. Η μέθοδος είναι επομένως καταρχήν ακατάλληλη για τις συνήθεις μετρολογικές εφαρμογές, όπου τα δεδομένα χαρακτηρίζονται από λιγότερο ή περισσότερο σημαντικές αβεβαιότητες, εκτός και αν υπάρχουν βάσιμοι λόγοι να θεωρηθούν οι αβεβαιότητες αυτές αμελητέες.

Συχνά επιλέγονται ως δείκτες αποτελεσματικότητας του μοντέλου συσχέτισης το αποκαλούμενο τυπικό σφάλμα $s_y$ (τετραγωνική ρίζα του πηλίκου του αθροίσματος των τετραγώνων των αποκλίσεων $e_i$ προς τους βαθμούς ελευθερίας $N$-$M$) και ο γνωστός σε όλους συντελεστής συσχέτισης $r$:

$$s_y = \sqrt{\sum e_i^2 / (N - M)} \ , \ r^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \qquad (3)$$

Από την παραπάνω σχέση φαίνεται ότι εάν το $r^2$ είναι μεγάλο (κοντά στην μονάδα), τότε το $s_y$ είναι μικρό, δηλαδή το μοντέλο συμπεριφέρεται ικανοποιητικά. Αυτό βέβαια ισχύει όταν δεν υπάρχουν σφάλματα στις πρωτογενείς μετρήσεις, σφάλματα τα οποία θα μπορούσαν να "αλλοιώσουν" την επεξηγηματική ικανότητα του μοντέλου. Πολύ συχνά χρησιμοποιείται εξάλλου ως αβεβαιότητα στις προβλέψεις του μοντέλου, όταν αυτό χρησιμοποιείται μετά την εύρεση των χαρακτηριστικών του συντελεστών, το τυπικό σφάλμα $s_y$, κυρίως γιατί υπολογίζεται αυτόματα από τα συνήθως χρησιμοποιούμενα λογισμικά στατιστικής ανάλυσης. Πρέπει να τονιστεί ότι η επιλογή αυτή είναι προβληματική και ασύμβατη με τις τρέχουσες αντιλήψεις για την αβεβαιότητα. Το τυπικό σφάλμα $s_y$, όπως και το τυπικό σφάλμα στις τιμές των παραμέτρων, δίνει μια μέση εικόνα για τις αποκλίσεις του μοντέλου από τις πειραματικά διαπιστωμένες τιμές του $y$, και όχι από τις ορθές τιμές του $y$, δεδομένου ότι για τον υπολογισμό του δεν πάρθηκαν υπόψη τα σφάλματα των πειραματικών παρατηρήσεων. Δεν επιτρέπει, κατά συνέπεια, να εκτιμηθούν τα όρια του εισαγόμενου σφάλματος από τη μελλοντική χρήση του μοντέλου συσχέτισης.

Τα μειονεκτήματα της κλασικής προσέγγισης των ελαχίστων τετραγώνων εστιάζονται κυρίως στο ότι κατά την αναζήτηση των συντελεστών του μοντέλου δεν συνυπολογίζονται οι αβεβαιότητες που χαρακτηρίζουν τις πειραματικές παρατηρήσεις και στο ότι δεν είναι δυνατή η αξιολόγηση της καταλληλότητας του μοντέλου στο να εξηγήσει τα πειραματικά δεδομένα.

Μια παραπλήσια προσέγγιση είναι αυτή των *Εκτιμητών Μεγίστης Πιθανόφάνειας* (*Maximum Likelihood Estimators*) η οποία, παρά το ότι κατά καιρούς έχει παρουσιαστεί ως διαφορετική, ανάγεται τελικά σε αυτή των ελαχίστων τετραγώνων [3, 4].

Η αρχή των Εκτιμητών Μεγίστης Πιθανότητας στηρίζεται στο ότι ορισμένα σετ τιμών των συντελεστών $a_j$ οδηγούν σε σημαντικές αποκλίσεις του μοντέλου από τα δεδομένα, ενώ άλλα σε μικρότερες. Από τη στιγμή που το «ορθό» μοντέλο είναι μόνο ένα, η ερώτηση *"με δεδομένες τις παρατηρήσεις ($x_{ki}$, $y_i$), ποια είναι η πιθανότητα να είναι σωστό ένα συγκεκριμένο σετ παραμέτρων $a_j$*" δεν είναι η πλέον κατάλληλη. Η ερώτηση αυτή μπορεί να αντιστραφεί και να τεθεί ως ακολούθως: "*με δεδομένο ένα συγκεκριμένο σετ*

*παραμέτρων $a_j$, ποια είναι η πιθανότητα εμφάνισης των παρατηρήσεων ($x_{ki}$, $y_i$), θεωρώντας ένα ορισμένο διάστημα εμπιστοσύνης $\Delta y$ για κάθε σημείο παρατήρησης*;" Με άλλα λόγια αναζητείται η πιθανότητα, ή καλύτερα η πιθανοφάνεια (*likelihood*) να εμφανιστούν οι παρατηρήσεις *($x_{ki}$, $y_i$)* για συγκεκριμένες τιμές των συντελεστών με τελικό στόχο να προσδιοριστεί το σετ εκείνο των συντελεστών $a_j$ για το οποίο μεγιστοποιείται η πιθανότητα αυτή.

Υποθέτουμε ότι για κάθε σημείο δεδομένων η πειραματική τιμή $y_i$ χαρακτηρίζεται από ένα σφάλμα το οποίο είναι ανεξάρτητο και ακολουθεί μια κανονική κατανομή γύρω από την θεωρούμενη ως αληθή τιμή $f(x_{1i}, x_{2i}, ..., x_{Ki})$ με τυπική απόκλιση $\sigma_i$. Η πιθανότητα εμφάνισης της τιμής $y_i$ μέσα στο διάστημα $\Delta y$ είναι επομένως:

$$\Delta y \, e^{-\frac{1}{2}\left(\frac{y_i - f(x_{1i}, x_{2i}, ..., x_{Ki})}{\sigma_i}\right)^2} \tag{4}$$

ενώ η πιθανότητα να συμβεί αυτό για όλα τα σημεία παρατήρησης ταυτόχρονα είναι το γινόμενο των επιμέρους πιθανοτήτων. Η μεγιστοποίηση επομένως της πιθανότητας αυτής περνάει από την ελαχιστοποίηση της ποσότητας:

$$\left[\sum_{i=1}^{N} \frac{\left[y_i - f(x_{1i}, x_{2i}, ..., x_{Ki})\right]^2}{2\sigma_i^2}\right] - N \log \Delta y \tag{5}$$

Δεδομένου ότι οι ποσότητες $N$ και $\Delta y$ είναι σταθερές και υποθέτοντας την ίδια τυπική απόκλιση $\sigma = \sigma_i$ για όλα τα σημεία, η σχέση (5) ανάγεται στη γνωστή σχέση των ελαχίστων τετραγώνων. Κατά συνέπεια, υποθέτοντας κανονική κατανομή (υπόθεση η οποία δεν είναι απαραίτητη για την κλασική μέθοδο των ελαχίστων τετραγώνων), οι εκτιμητές ελαχίστων τετραγώνων και μεγίστης πιθανότητας ταυτίζονται στην πράξη, με τους δεύτερους όμως να επιτρέπουν την εκτίμηση των αβεβαιοτήτων στους συντελεστές, υποθέτοντας όπως προαναφέρθηκε την ίδια τυπική απόκλιση για όλα τα $y$ [5].

## 3. Συσχέτιση δεδομένων σε συνθήκες αβεβαιότητας

Μια άλλη πολύ πιο ενδιαφέρουσα εναλλακτική δυνατότητα συνίσταται στην χρησιμοποίηση της μεθόδου των ***weighted least squares (WLS)*** η οποία, στη βάση των μετρούμενων τιμών αλλά και της αβεβαιότητας τους, υπολογίζει όχι μόνο τους συντελεστές του μοντέλου αλλά και την αβεβαιότητα τους.

Στην πραγματική ζωή όλες οι παρατηρήσεις χαρακτηρίζονται από αβεβαιότητες οι οποίες διαφέρουν κατά κανόνα από το ένα σημείο παρατήρησης στο άλλο. Η συσχέτιση των πρωτογενών δεδομένων για τον προσδιορισμό των συντελεστών του μοντέλου οφείλει να λάβει υπόψη της τις αβεβαιότητες αυτές, όχι μόνο για να είναι αντικειμενική η εκτίμηση των ίδιων των συντελεστών, αλλά και για να είναι ρεαλιστική η αποτίμηση των αβεβαιοτήτων που θα χαρακτηρίζουν τη μελλοντική χρήση του μοντέλου. Για παράδειγμα, ένα σημείο παρατήρησης κακής μετρολογικής ποιότητας (χαρακτηριζόμενο από μεγάλες αβεβαιότητες), είναι λογικό να συμβάλει λιγότερο στη διαμόρφωση των συντελεστών του μοντέλου συσχέτισης, συγκριτικά με ένα άλλο σημείο υψηλότερης μετρολογικής ποιότητας.

Μια φυσική προσέγγιση για την αντιμετώπιση του προβλήματος της συσχέτισης σε συνθήκες αβεβαιότητας συνίσταται στην θεώρηση της κάθε παρατήρησης ως πηγή νέας γνώσης για τη συμπεριφορά των μεταβλητών και του ίδιου του μοντέλου που τις συνδέει. Τοποθετημένου λοιπόν του προβλήματος, στα πλαίσια μιας Bayesian συλλογιστικής, η συνάρτηση βελτιστοποίησης μπορεί να οριστεί ως ακολούθως [4, 6]:

$$\chi^2 = \sum_{i=1}^{N} \left[ \frac{y_i - f(x_{1i}, x_{2i}, ..., x_{Ki})}{\sigma_i} \right]^2 \qquad (6)$$

Η ποσότητα $\sigma_i^2$ αντιπροσωπεύει τη μεταβλητότητα για κάθε σημείο παρατήρησης και εξαρτάται από τις πειραματικές αβεβαιότητες που χαρακτηρίζουν την κάθε παρατήρηση, τόσο στις τιμές της εξαρτημένης όσο και σε αυτές των ανεξάρτητων μεταβλητών. Με βάση την εξίσωση (6), η ποσότητα $\chi^2$ δίνει ουσιαστικά μια ιδέα για την σχέση ανάμεσα στην απόκλιση μεταξύ μοντέλου και πειραματικών δεδομένων από τη μια, και αβεβαιότητας των μετρήσεων από την άλλη. Αποτελεί με άλλα λόγια ένα δείκτη του κατά πόσο οι παρατηρούμενες αποκλίσεις μπορούν να εξηγηθούν από τα σφάλματα των μετρήσεων. Επιπλέον, η προσέγγιση αυτή οδηγεί στην εύρεση εκείνου του σετ συντελεστών $a_1, a_2, ...a_M$ του μοντέλου συσχέτισης που ελαχιστοποιούν τις αποκλίσεις όχι ως απόλυτες τιμές, αλλά σε σχέση με την αβεβαιότητά τους.
Η μεταβλητότητα $\sigma_i^2$ της κάθε απόκλισης υπολογίζεται από τη σχέση:

$$\sigma_i^2 = Var\left(y_i - f(x_{1i}, x_{2i}, ..., x_{Ki})\right) \qquad (7)$$

Να σημειωθεί ότι, παρά τη φαινομενική τους ομοιότητα, οι σχέσεις (5) και (6) διαφέρουν κατά το ότι στην πρώτη θεωρείται ότι τα πειραματικά σφάλματα αφορούν μόνο την εξαρτημένη μεταβλητή, ενώ στη δεύτερη λαμβάνονται υπόψη και τα σφάλματα στις ανεξάρτητες μεταβλητές.
Από τις σχέσεις (1) και (7) και θεωρώντας ανεξάρτητες μεταξύ τους τις μεταβλητές $x_1$, $x_2, ...x_K$, προκύπτει:

$$\sigma_i^2 = u_{yi}^2 + \sum \left( \frac{\partial f}{\partial x_k} \right)^2 u_{Xki}^2 = u_{yi}^2 + \sum_{j=1}^{M} a_j^2 \sum_{k=1}^{K} \left( \frac{\partial f_j}{\partial x_k} \right)^2 u_{Xki}^2 \qquad (8)$$

όπου $u_{yi}$ και $u_{Xki}$ είναι οι τυπικές αβεβαιότητες των $y_i$ και $x_{ki}$ αντίστοιχα.

Από τις σχέσεις (6) έως (8) γίνεται φανερό ότι ο υπολογισμός των συντελεστών $a_1$, $a_2, ...a_M$ μέσω της ελαχιστοποίησης της σχέσης (6) προϋποθέτει τη γνώση της μεταβλητότητας $\sigma_i^2$ και άρα των ίδιων των συντελεστών, πράγμα που δυσκολεύει την διαδικασία επίλυσης. Μία πρακτική λύση συνίσταται στην επίλυση του κλασικού προβλήματος ελαχίστων τετραγώνων, δηλαδή χωρίς να ληφθούν υπόψη οι αβεβαιότητες, και ο υπολογισμός κάποιων αρχικών τιμών των συντελεστών $a_j$. Οι αρχικές αυτές τιμές χρησιμοποιούνται για τον υπολογισμό των $\sigma_i$ με τη βοήθεια της σχέσης (8). Στη συνέχεια ελαχιστοποιείται η ποσότητα $\chi^2$ της εξίσωσης (6) και προσδιορίζονται νέες τιμές των $a_j$. Οι νέες αυτές τιμές χρησιμοποιούνται για τον υπολογισμό των $\sigma_i$ και την εκ νέου επίλυση της εξίσωσης και η διαδικασία συνεχίζεται επαναληπτικά μέχρι την οριστική σύγκλιση.

Η ελαχιστοποίηση είναι δυνατή χρησιμοποιώντας τις συνήθεις μεθόδους μη γραμμικής ελαχιστοποίησης. Σε μορφή διανυσμάτων και πινάκων, εάν Α είναι ένας πίνακας του οποίου τα *NxM* στοιχεία ορίζονται από τις στοιχειώδεις συναρτήσεις στα σημεία ($y_i$; $x_{1i}$, $x_{2i}$,...$x_{Ki}$) και από την αντίστοιχη τυπική απόκλιση $\sigma_i$ [4,8]:

$$A_{ij} = \frac{f_j}{\sigma_i}, \qquad \mathrm{A}=\begin{vmatrix} \frac{f_1}{\sigma_1} & \frac{f_2}{\sigma_1} & \cdots & \frac{f_M}{\sigma_1} \\ \frac{f_1}{\sigma_2} & \frac{f_2}{\sigma_2} & \cdots & \frac{f_M}{\sigma_1} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \frac{f_1}{\sigma_N} & \frac{f_2}{\sigma_N} & \cdots & \frac{f_M}{\sigma_N} \end{vmatrix} \tag{9}$$

Έστω επίσης το άνυσμα Β μήκους Ν:

$$b_i = \frac{y_i}{\sigma_i}, \qquad B = \begin{vmatrix} y_1 / \sigma_1 \\ \cdot \\ \cdot \\ y_N / \sigma_N \end{vmatrix} \tag{10}$$

Η εξίσωση του γενικού προβλήματος των ελαχίστων τετραγώνων (*normal equation of the least square problem*) μπορεί τότε να διατυπωθεί ως εξής:

$$\left(A^T \cdot A\right) \cdot INV(C) = A^T \cdot B \tag{11}$$

Τα στοιχεία της C είναι οι συντελεστές $a_j$ του μοντέλου. Μπορεί επίσης να οριστεί ο πίνακας $Z = INV(A^T \cdot A)$ του οποίου τα διαγώνια στοιχεία $z_{jj}$ είναι οι μεταβλητότητες $u_{aj}^2$ (τετράγωνο της τυπικής αβεβαιότητας) των συντελεστών $a_j$, ενώ τα μη-διαγώνια στοιχεία της $z_{jk}$ είναι οι συμμεταβλητότητές τους $Cov(a_j,a_k)$ (covariances):

$$u_{aj}^2 = z_{jj}, \quad Cov(a_j, a_k) = z_{jk} \tag{12}$$

Η επίλυση της εξίσωσης (11) μπορεί να γίνει με μια κλασική μέθοδο (LU decomposition, Gauss-Jordan elimination κλπ). Είναι επίσης δυνατή με χρήση ευρέως διαδεδομένων λογισμικών όπως το Excel, που υποστηρίζουν υπολογισμούς σε πίνακες.

Πρέπει να σημειωθεί ότι οι τιμές των συντελεστών $a_1, a_2,...a_M$ που υπολογίζονται με τη μέθοδο των σταθμισμένων ελαχίστων τετραγώνων είναι, εν γένει, διαφορετικές από αυτές που θα προέκυπταν εάν είχε εφαρμοστεί η μέθοδος των απλών ελαχίστων τετραγώνων, γεγονός που αποτελεί ένα επιπλέον επιχείρημα υπέρ της χρήσης των πρώτων.

Από τη στιγμή που υπολογίστηκαν οι συντελεστές $a_1, a_2,...a_M$, οι τυπικές τους αβεβαιότητες $u_{aj}$ και οι συμμεταβλητότητες $Cov(a_j,a_k)$, ο υπολογισμός της αβεβαιότητας στο αποτέλεσμα *y* του μοντέλου, όταν αυτό χρησιμοποιηθεί στη συνέχεια για δεδομένες τιμές $x_1, x_2,...x_K$ των ανεξάρτητων μεταβλητών, γίνεται εύκολα εφαρμόζοντας το νόμο διάδοσης των αβεβαιοτήτων [1, 5, 7]:

$$u_y^2 = \sum_{j=1}^{M} \left( \frac{\partial f}{\partial a_j} \right)^2 u_{aj}^2 + 2 \sum_{k=1}^{M-1} \sum_{l=k+1}^{M} \frac{\partial f}{\partial a_k} \frac{\partial f}{\partial a_l} Cov(a_k, a_l) \qquad (13)$$

Στη συνέχεια είναι δυνατόν να προσδιοριστεί κατά τα συνήθη ένα διάστημα εμπιστοσύνης για το *y*, για μια δεδομένη πιθανότητα κάλυψης, θεωρώντας όμως *N-M* βαθμούς ελευθερίας, δεδομένου ότι οι συντελεστές του μοντέλου δεν είναι όλοι στατιστικά ανεξάρτητοι.

Μια άλλη ενδιαφέρουσα πτυχή της προσέγγισης αυτής είναι η δυνατότητα να αξιολογηθεί η καταλληλότητα του δεδομένου μοντέλου, δηλαδή το κατά πόσο το μοντέλο αυτό «εξηγεί» τις πειραματικές παρατηρήσεις. Πιο συγκεκριμένα, είναι γνωστό ότι η στατιστική $\chi^2$ έχει μέσο όρο το $\nu$ και τυπική απόκλιση το $(2\nu)^{0,5}$ και τείνει ασυμπτωτικά σε κανονική κατανομή όταν το $\nu$ παίρνει μεγάλες τιμές [4]. Η πιθανότητα *Q* να εμφανιστεί κατά τύχη μια τιμή του $\chi^2$ τόσο χαμηλή όσο η ελάχιστη τιμή $\chi^2_{\min}$ (που οδηγεί στην εύρεση των συντελεστών του μοντέλου συσχέτισης), ακόμα και για ένα σωστό μοντέλο, δίδεται από την σχέση:

$$Q\left(\nu, \chi^2_{\min}\right) = gammq\left(\frac{\nu}{2}, \frac{\chi^2_{\min}}{2}\right) \qquad (14)$$

όπου *gammq(a,x)* είναι η ατελής συνάρτηση gamma:

$$gammq(a,x) = \frac{1}{\Gamma(a)} \int_x^{\infty} e^{-t} t^{a-1} dt \qquad (15)$$

Η πιθανότητα αυτή δίδει μια ποσοτική ένδειξη της καταλληλότητας του μοντέλου. Εάν αυτή η πιθανότητα είναι σημαντική, μεγαλύτερη από 0,1 για παράδειγμα, τότε το μοντέλο μπορεί να θεωρηθεί αξιόπιστο. Εάν η πιθανότητα *Q* είναι πολύ μικρή για ένα συγκεκριμένο σετ δεδομένων, είναι μάλλον απίθανο οι αποκλίσεις να είναι τυχαίες και μπορεί να οφείλονται είτε στην ακαταλληλότητα του μοντέλου, είτε στην κακή εκτίμηση του μεγέθους του σφάλματος $\sigma_i$ το οποίο είναι μεγαλύτερο από αυτό που υπολογίστηκε, είτε τέλος στο ότι τα σφάλματα δεν ακολουθούν κανονική κατανομή. Συμβατικά θεωρείται ότι όταν η πιθανότητα *Q* είναι μικρή αλλά μεγαλύτερη από 0.001, τότε η συσχέτιση μπορεί να γίνει αποδεκτή εάν τα σφάλματα δεν ακολουθούν κανονική κατανομή ή έχουν υποτιμηθεί. Εάν όμως το *Q* είναι μικρότερο από 0.001, τότε το μοντέλο μπορεί να αμφισβητηθεί.

## 4. Εφαρμογή στην εύρεση καμπύλης διακρίβωσης

Μια συνήθης περίπτωση διακρίβωσης είναι όταν συγκρίνονται οι ενδείξεις $d_i$ του υπό διακρίβωση οργάνου με τις ενδείξεις $r_i$ ενός οργάνου αναφοράς σε $i=1...N$ σημεία διακρίβωσης. Αναζητείται, με την μέθοδο των σταθμισμένων ελαχίστων τετραγώνων, μια γραμμική καμπύλη διακρίβωσης της μορφής η οποία μπορεί να χρησιμοποιηθεί για τη διόρθωση των μελλοντικών ενδείξεων του οργάνου, μετά τη διακρίβωση του [9]:

$$r = a_0 + a_1 d \qquad (16)$$

Στόχος είναι, με την ελαχιστοποίησης του $\chi^2$, να υπολογιστούν όχι μόνο οι τιμές των συντελεστών $a_0$ και $a_1$ αλλά και οι αβεβαιότητα που τις χαρακτηρίζει. Η εφαρμογή των σχέσεων (6) έως (8) με $M=2$ και $N=9$ οδηγεί στη σχέση:

$$\chi^2 = \sum_{i=1}^{N} \frac{(r_i - a_0 - a_1 d_i)^2}{u_{ri}^2 + a_1^2 u_{di}^2} \tag{17}$$

Η επίλυση του προβλήματος ελαχιστοποίησης με τη μέθοδο που εκτέθηκε παραπάνω, επιτρέπει τον υπολογισμό των κατάλληλων τιμών των συντελεστών $a_0$ και $a_1$, των αντίστοιχων αβεβαιοτήτων $u_{a0}$ και $u_{a1}$, καθώς και της συμμεταβλητότητάς τους $Cov(a_0,a_1)$. Στη συνέχεια είναι εύκολο να υπολογιστεί η αβεβαιότητα $u_E$ στις μελλοντικές διορθωμένες τιμές $E$ που προκύπτουν από τις ενδείξεις $e$ του διακριβωμένου οργάνου:

$$E=a_0+a_1 e \tag{18}$$

$$u_E = \sqrt{u_{a0}^2 + e^2 u_{a1}^2 + 2eCov(a_0,a_1)} \tag{19}$$

## 5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η συσχέτιση δεδομένων στα πλαίσια των μετρολογικών εφαρμογών δεν μπορεί να θεωρηθεί πλήρης εάν δεν συνυπολογίζει την ποιότητα των συσχετιζόμενων δεδομένων, δηλαδή την αβεβαιότητα που τα χαρακτηρίζει. Η κλασική προσέγγιση των απλών ελαχίστων τετραγώνων είναι από την άποψη αυτή ανεπαρκής, στο βαθμό που αδυνατεί να οδηγήσει σε μια αποτίμηση της ποιότητας του ίδιου του μοντέλου συσχέτισης αλλά και των σφαλμάτων που εισάγονται από την χρήση του.
Η μέθοδος των σταθμισμένων ελαχίστων τετραγώνων επιτρέπει, αντίθετα, τον προσδιορισμό ενός ρεαλιστικού μοντέλου, την αποτίμηση της καταλληλότητάς του, αλλά και την βελτίωση της αποτελεσματικότητάς του, στο βαθμό που επιτρέπει την έγκυρη εκτίμηση αβεβαιοτήτων που θα χαρακτηρίζουν τις τιμές που το μοντέλο αυτό θα παράγει μετά την ολοκλήρωση της διαδικασίας συσχέτισης.

## ΒΙΒΛΙΟΓΡΑΦΙΑ


[1] ISO (1995), *Guide to the expression of uncertainty in measurements*, ISO 2nd ed., Switzerland.
[2] Bajpai A.C., Mustoe L.R. and Walker D. (1977), *Advanced engineering mathematics*, John Wiley, New York.
[3] Dietrich C.F. (1991), *Uncertainty, calibration and probability*, 2nd ed, Adam-Hilger, Bristol.
[4] Press W., Teukolsky S.A., Vetterling W.T. and Flannery B.P. (1996), *Numerical recipes*, 2nd edn., Cambridge University Press, Oxford.
[5] Μαθιουλάκης Ε. (2004), *Μέτρηση, ποιότητα μέτρησης και αβεβαιότητα*, Έκδοση HellasLab, Αθήνα
[6] Lira I., *Evaluating the measurement uncertainty,* IoP ed., 2002.

[7] Reis M. S. and Saraiva P. M., *Integration of data uncertainty in linear regression and process optimization,* AIChE Journal, Vol. 51, No. 11 (2005), 3007-3018.
[8] Brown P. G., (1993), *Measurement, regression and calibration*, Clarendon Press, Oxford.
[9] Mathioulakis E. and Belessiotis V., *Uncertainty and traceability in calibration by comparison*, Meas. Sci. Technol., 11 (2000), 771-775.