Journal of Geodetic Science
Research Article Open Access

Tik Mok and H. Bâki Iz*

Vector regression introduced

Abstract: This study formulates regression of vector data that will enable statistical analysis of various geodetic phenomena such as, polar motion, ocean currents, typhoon/hurricane tracking, crustal deformations, and precursory earthquake signals. The observed vector variable of an event (dependent vector variable) is expressed as a function of a number of hypothesized phenomena *realized also as vector variables* (independent vector variables) and/or scalar variables that are likely to impact the dependent vector variable. The proposed representation has the *unique* property of solving the coefficients of independent vector variables (explanatory variables) also *as vectors*, hence it supersedes multivariate multiple regression models, in which *the unknown coefficients are scalar quantities*. For the solution, complex numbers are used to represent vector information, and the method of least squares is deployed to estimate the vector model parameters after transforming the complex vector regression model into a real vector regression model through isomorphism. Various operational statistics for testing the predictive significance of the estimated vector parameter coefficients are also derived. A simple numerical example demonstrates the use of the proposed vector regression analysis in modeling typhoon paths.

Keywords: complex least squares adjustment, vector data, vector regression

DOI 10.2478/jogs-2014-0009
Received January 12, 2014 ; accepted April 4, 2014.

1 Introduction

Recent proliferation of rapid acquisition of position information generates vast amounts of vector data. This study formulates regression of such vector data that will enable exploratory analysis as well as explanatory representation of various geodetic phenomena including polar motion, crustal deformation, precursory earthquake signals, or any other kinematic event under the umbrella of geodetic/geospatial information, relating them to a number of underlying effects that can be expressed as vector and scalar information in combination. In this framework, vector regression models can be semi-empirical or empirical and they can be used to model statistical relationships i.e., correlations among dependent and, independent vectors in combination with scalar parameters akin to well-known regression analysis.

The vector regression approach proposed in this study differs from the bidimensional regression formulation by Tobler (1994) in which the regression coefficients constitute a spatially varying, but coordinate invariant, second-order tensor field defined by the matrix of partial derivatives of the transformation, which is purely geometric. The approach in this study also differs conceptually from the mathematical model proposed by Schneider (1982), which exploits the geometric transformational properties of complex numbers in describing two dimensional strain.

Vector regression model also differs from multivariate regression models because, the observed vector variable of an event (dependent vector variable) is expressed as a function of a number of hypothesized phenomena *realized also as vector variables* (independent vector variables) and/or scalar variables that are likely to impact the dependent vector variable. Hence, vector regression models supersede multivariate multiple regression models in which the unknown coefficients of the regressors (independent variables) are scalar quantities as opposed to unknown vector correlates in vector regression models.

In vector regression, it is possible, for instance, to establish a statistical model in which post-earthquake displacements, as vector dependent variables having a direction and magnitude, are related to the propagation direction and magnitude of epicenters of micro seismic events also represented as vectors observed prior to earthquakes (or foreshocks magnitudes and their fault plane orientations). Such a model can be used to investigate potential statistical relationships, which are the unknown vector coefficients, operating as vector correlates between the two events. Such a model can also be extended to include, co-

Tik Mok: Dept. of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China
***Corresponding Author: H. Bâki Iz:** Dept. of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China, E-mail: lshbiz@polyu.edu.hk

variates such as displacements observed in the surrounding areas as a result of prior earthquakes having a lagged impact in triggering the recent event (or and the *a posteriori* vector crustal displacements together with micro seismicity. Currently such investigations cannot be carried out simultaneously in any statistical model including multivariate models. A vector regression model, having statistically significant predictive power following a proper statistical hypothesis testing, can potentially be used as an explanatory tool for earthquake prediction studies.

In this study, first, a number of fundamental definitions of the complex algebra that are used in the derivations are presented. Complex numbers, complex vectors, and complex matrices are used to establish a vector regression model in the following section. The least squares method is implemented to estimate the model parameters after transforming the complex vector model into a real vector mode through isomorphism. In addition to the solution formulation, fundamental statistics such as total sum of squares, sum of squares errors as well as methods for null-hypothesis testing of the model and vector parameters of the vector regression model are derived. A simple model for typhoon path prediction demonstrates numerically the use of the proposed vector regression formulation and its solution before the conclusion.

2 Definitions for complex matrix algebra

The *modus operandi* of the proposed vector regression model involves manipulating complex numbers, complex vectors, and complex matrices. Complex vectors and complex matrices are vectors and matrices whose components can be complex numbers. They were introduced by J. W. Gibbs during 1880s at widely circulated pamphlet Elements of Vector Analysis (Wilson, 1901).

The relevant definitions are as follows (Anderson et al., 1995):

Definition 2.1. A complex number c is defined as $c := a + ib \in \mathbb{C}$, where i is the imaginary number, $\mathbb{C}$ is a complex number space, and a and b belong to a real number space $\mathbb{R}$.

Definition 2.2. Let A and B represent $n \times 1$ vectors with $A, B \in \mathbb{R}^{n\times 1}$ where $\mathbb{R}^{n\times 1}$ is a $n \times 1$ real number space. Then, $C := A + iB \in \mathbb{C}^{n\times 1}$, which is a complex number space with dimension $n \times 1$. C is a $n \times 1$ complex vector whose elements are defined as complex numbers $c_j = a_j + ib_j$ with $j = 1, 2, \dots, n$.

Definition 2.3. Let $\mathbf{A}$ and $\mathbf{B}$ represent $n \times p$ matrices with $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times p}$ where $\mathbb{R}^{n\times p}$ is a $n \times p$ real number space. Then, $\mathbf{C} = \mathbf{A} + i\,\mathbf{B} \in \mathbb{C}^{n\times p}$, which is a complex number space with dimension $n \times p$. $\mathbf{C}$ is an $n \times p$ complex matrix whose elements are defined as complex numbers $c_{jk} = a_{jk} + ib_{jk}$ with $j = 1, 2, \dots, n$ and $k = 1, 2, \dots, p$.

Definition 2.4. For a complex number $c := a + ib$, its real isomorphism, denoted by [.], is given by $[c] = \begin{pmatrix} a \\ b \end{pmatrix}$. Note that the role of isomorphism is to convert the complex number from $\mathbb{C}$ to $\mathbb{R}^2$, where $\mathbb{C}$ is a complex number space and $\mathbb{R}^2$ is a real number space with twice the dimension of $\mathbb{C}$.

Definition 2.5. To perform a complex multiplication in isomorphism form, a particular 2×2 real matrix $\{c\} \in \mathbb{R}^{2\times 2}$, is defined as $\{c\} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$.

Example 2.6. Let $c = a + ib$, $d = e + ih$, where c and d are complex numbers that belong to $\mathbb{C}$ and where a, b, e and h are real numbers that belong to $\mathbb{R}$. Then, $c \cdot d = (ae - bh) + i\,(ah + be) \Leftrightarrow\Leftrightarrow [c]\cdot[d] = \{c\}\cdot[d] = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \cdot \begin{pmatrix} e \\ h \end{pmatrix} = \begin{pmatrix} ae - bh \\ ah + be \end{pmatrix}$.

Let $\mathbf{A}$ and $\mathbf{B}$ represent $n \times p$ matrices, i.e. $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times p}$ and $\mathbf{C}$ is an $n \times p$ complex matrix. Then the $2n \times 2p$ real matrix $\{\mathbf{C}\} \in \mathbb{R}^{2n\times 2p}$ is defined as $\{\mathbf{C}\} := \begin{pmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{pmatrix}$.

Definition 2.7. Every element in a conjugate matrix are conjugate to the original complex elements, i.e. for a complex number $c = a + ib$, the complex conjugate number is $c^\star = a - ib$. $\mathbf{C}^*$ denotes the $p \times n$ complex conjugate matrix. Note that by definition, complex conjugate matrix is the transpose of the complex matrix $\mathbf{C}$ in which all the complex numbers are replaced by their complex conjugate numbers. The asterisk operates like matrix transpose in real space, i.e. $(\mathbf{CD})^* = \mathbf{D}^*\mathbf{C}^*$.

Definition 2.8. A complex random variable x is defined as $x := u + iv$, where u and v are real random numbers, whose range is (a subset of) the real-valued Euclidean space.

3 Complex least squares adjustment

The linear regression model for the vector data reads as:

$$Y = \mathbf{X}\beta + \varepsilon \tag{1}$$

where the Y vector is a complex vector of the dependent variable that belongs to a $n \times 1$ vector space, β is a $k \times 1$ complex coefficient vector, ε is a $n\times1$ complex error vector, and $\mathbf{X}$ is a $n \times k$ complex design matrix and assumed to be deterministic in nature.

It is assumed that the real and imaginary part of the complex error vector ε are independently and identically distributed, i.e.,

$$\begin{aligned}&E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) := E\left(\varepsilon_i{}^{*}\varepsilon_i\right) = \sigma^2 \text{ for all } i,\\ &\text{and } Cov\left(\varepsilon_i, \varepsilon_j\right) = 0 \text{ for } i \neq j \Rightarrow E(\varepsilon\varepsilon^{*}) = \sigma^2 I\end{aligned} \tag{2}$$

where $\varepsilon_i s$ are the components of the complex error vector ε.

In tandem, the complex residual vector $\hat{\varepsilon}$ is defined as:

$$\hat{\varepsilon} = Y - \mathbf{X}\hat{\beta} \tag{3}$$

In which $\hat{\beta}$ is an estimate of the unknown $k \times 1$ complex coefficient vector β (vector correlates). Eq. (3) can be expressed in a partitioned format using isomorphism following Definition 2.4 and Definition 2.5 as follows:

$$\left[\hat{\varepsilon}\right] = [Y] - \{\mathbf{X}\}\left[\hat{\beta}\right] = \begin{pmatrix} Y_u \\ Y_v \end{pmatrix} - \begin{pmatrix} \mathbf{X}_u & -\mathbf{X}_v \\ \mathbf{X}_v & \mathbf{X}_u \end{pmatrix}\begin{pmatrix} \hat{\beta}_u \\ \hat{\beta}_v \end{pmatrix}. \tag{4}$$

In this expression Y_u , Y_v are the real and the imaginary parts of Y. $\mathbf{X}_u$ and $\mathbf{X}_v$ are the real and the imaginary parts of $\mathbf{X}$, and $\hat{\beta}_u$, $\hat{\beta}_v$ are the real and the imaginary parts of $\hat{\beta}$.

Similar to ordinary regression with real numbers, the *residual sum of squares, RSS*, is defined as:

$$RSS := \left(Y - \mathbf{X}\hat{\beta}\right)^{*}\left(Y - \mathbf{X}\hat{\beta}\right) \tag{5}$$

Considering Eq. (4), Eq. (5) can now be expressed as:

$$RSS = \begin{pmatrix} Y_u - \mathbf{X}_u\hat{\beta}_u + \mathbf{X}_v\hat{\beta}_v \\ Y_v - \mathbf{X}_v\hat{\beta}_u - \mathbf{X}_u\hat{\beta}_v \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} Y_u - \mathbf{X}_u\hat{\beta}_u + \mathbf{X}_v\hat{\beta}_v \\ Y_v - \mathbf{X}_v\hat{\beta}_u - \mathbf{X}_u\hat{\beta}_v \end{pmatrix} \tag{6}$$

or equivalently:

$$\begin{aligned}RSS &= Y_u{}^{\mathrm{T}}Y_u - 2Y_u{}^{\mathrm{T}}\mathbf{X}_u\hat{\beta}_u + 2Y_u{}^{\mathrm{T}}\mathbf{X}_v\hat{\beta}_v + \hat{\beta}_u^{\mathrm{T}}\mathbf{X}_u{}^{\mathrm{T}}\mathbf{X}_u\hat{\beta}_u\\ &+ \hat{\beta}_v^{\mathrm{T}}\mathbf{X}_v{}^{\mathrm{T}}\mathbf{X}_v\hat{\beta}_v + +Y_v{}^{\mathrm{T}}Y_v - 2Y_v{}^{\mathrm{T}}\mathbf{X}_v\hat{\beta}_u - 2Y_v{}^{\mathrm{T}}\mathbf{X}_u\hat{\beta}_v\\ &+ \hat{\beta}_u^{\mathrm{T}}\mathbf{X}_v{}^{\mathrm{T}}\mathbf{X}_v\hat{\beta}_u + \hat{\beta}_v^{\mathrm{T}}\mathbf{X}_u{}^{\mathrm{T}}\mathbf{X}_u\hat{\beta}_v\end{aligned} \tag{7}$$

The differential of Eq. (7) in light of $\left[\hat{\beta}\right] = \begin{pmatrix} \hat{\beta}_u & \hat{\beta}_v \end{pmatrix}^{T}$ by (Definition 2.4) gives:

$$\frac{\partial RSS}{\partial[\hat{\beta}]} = \begin{pmatrix} -2Y_u{}^{\mathrm{T}}\mathbf{X}_u - 2Y_v{}^{\mathrm{T}}\mathbf{X}_v + 2\hat{\beta}_u^{\mathrm{T}}\mathbf{X}_u{}^{\mathrm{T}}\mathbf{X}_u + 2\hat{\beta}_u^{\mathrm{T}}\mathbf{X}_v{}^{\mathrm{T}}\mathbf{X}_v \\ +2Y_u{}^{\mathrm{T}}\mathbf{X}_v - 2Y_v{}^{\mathrm{T}}\mathbf{X}_u + 2\hat{\beta}_v^{\mathrm{T}}\mathbf{X}_v{}^{\mathrm{T}}\mathbf{X}_v + 2\hat{\beta}_v^{\mathrm{T}}\mathbf{X}_u{}^{\mathrm{T}}\mathbf{X}_u \end{pmatrix} \tag{8}$$

Setting Eq. (8) equal to zero:

$$\begin{pmatrix} -2Y_u{}^{\mathrm{T}}\mathbf{X}_u - 2Y_v{}^{\mathrm{T}}\mathbf{X}_v + 2\hat{\beta}_u^{\mathrm{T}}\mathbf{X}_u{}^{\mathrm{T}}\mathbf{X}_u + 2\hat{\beta}_u^{\mathrm{T}}\mathbf{X}_v{}^{\mathrm{T}}\mathbf{X}_v \\ +2Y_u{}^{\mathrm{T}}\mathbf{X}_v - 2Y_v{}^{\mathrm{T}}\mathbf{X}_u + 2\hat{\beta}_v^{\mathrm{T}}\mathbf{X}_v{}^{\mathrm{T}}\mathbf{X}_v + 2\hat{\beta}_v^{\mathrm{T}}\mathbf{X}_u{}^{\mathrm{T}}\mathbf{X}_u \end{pmatrix} = \\ = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{9}$$

the following expression is obtained:

$$\begin{pmatrix} 2Y_u{}^{\mathrm{T}}\mathbf{X}_u + 2Y_v{}^{\mathrm{T}}\mathbf{X}_v \\ 2Y_v{}^{\mathrm{T}}\mathbf{X}_u - 2Y_u{}^{\mathrm{T}}\mathbf{X}_v \end{pmatrix} = \begin{pmatrix} 2\hat{\beta}_u{}^{\mathrm{T}}\mathbf{X}_u{}^{\mathrm{T}}\mathbf{X}_u + 2\hat{\beta}_u^{\mathrm{T}}\mathbf{X}_v{}^{\mathrm{T}}\mathbf{X}_v \\ 2\hat{\beta}_v^{\mathrm{T}}\mathbf{X}_v{}^{\mathrm{T}}\mathbf{X}_v + 2\hat{\beta}_v^{\mathrm{T}}\mathbf{X}_u{}^{\mathrm{T}}\mathbf{X}_u \end{pmatrix}. \tag{10}$$

Eq. (10) can be rearranged as:

$$\begin{pmatrix} Y_u \\ Y_v \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} \mathbf{X}_u & -\mathbf{X}_v \\ \mathbf{X}_v & \mathbf{X}_u \end{pmatrix} = \\ = \begin{pmatrix} \hat{\beta}_u \\ \hat{\beta}_v \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} \mathbf{X}_u{}^{\mathrm{T}}\mathbf{X}_u + \mathbf{X}_v{}^{\mathrm{T}}\mathbf{X}_v & 0 \\ 0 & \mathbf{X}_u{}^{\mathrm{T}}\mathbf{X}_u + \mathbf{X}_v{}^{\mathrm{T}}\mathbf{X}_v \end{pmatrix}, \tag{11}$$

followed by:

$$\begin{pmatrix} \mathbf{X}_u & -\mathbf{X}_v \\ \mathbf{X}_v & \mathbf{X}_u \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} Y_u \\ Y_v \end{pmatrix} = \begin{pmatrix} \mathbf{X}^{*}\mathbf{X} & 0 \\ 0 & \mathbf{X}^{*}\mathbf{X} \end{pmatrix}\begin{pmatrix} \hat{\beta}_u \\ \hat{\beta}_v \end{pmatrix} \Rightarrow \\ \Rightarrow \{\mathbf{X}\}^{\mathrm{T}}[Y] = \left\{\mathbf{X}^{*}\mathbf{X}\right\}\left[\hat{\beta}\right]. \tag{12}$$

Transformation of Eq. (12) into complex number form gives:

$$\hat{\beta} = \left(\mathbf{X}^{*}\mathbf{X}\right)^{-1}\mathbf{X}^{*}Y. \tag{13}$$

The variance/covariance matrix of the vector estimates, denoted by $cov\left(\hat{\beta}\right)$, can now be obtained utilizing model (1) and Eq. (13), which leads to:

$$\hat{\beta} = \beta + \left(\mathbf{X}^{*}\mathbf{X}\right)^{-1}\mathbf{X}^{*}\varepsilon \tag{14}$$

Considering the assumptions about the complex error vector properties given by Eq. (2), and the fundamental definition of the $cov\left(\hat{\beta}\right)$ given below

$$\mathrm{E}\left(\hat{\beta}\right) = \beta \tag{15}$$

$$cov\left(\hat{\beta}\right) := \mathrm{E}\left(\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\star}\right) \tag{16}$$

and substituting Eq. (14) into Eq. (16) results in:

$$\begin{aligned} cov\left(\hat{\beta}\right) &= \mathrm{E}\left(\left(\mathbf{X}^{\star}\mathbf{X}\right)^{-1}\mathbf{X}^{\star}\varepsilon\varepsilon^{\star}\mathbf{X}\left(\mathbf{X}^{\star}\mathbf{X}\right)^{-1}\right) \\ &= \left(\mathbf{X}^{\star}\mathbf{X}\right)^{-1}\mathbf{X}^{\star}\mathrm{E}\left(\varepsilon\varepsilon^{\star}\right)\mathbf{X}\left(\mathbf{X}^{\star}\mathbf{X}\right)^{-1} \\ &= \mathrm{E}\left(\varepsilon\varepsilon^{\star}\right)\left(\mathbf{X}^{\star}\mathbf{X}\right)^{-1}. \end{aligned} \tag{17}$$

The above expression, using Definition 2.7 in light of Eq. (2), reduces to,

$$cov\left(\hat{\beta}\right) = \sigma^2\left(\mathbf{X}^{\star}\mathbf{X}\right)^{-1}. \tag{18}$$

A number of solution statistics akin to ordinary regression can now be derived for the vector regression. Let the Sum of Squares, *SS*, of the predicted (adjusted) vector observations be denoted by:

$$SS := \hat{Y}^{\star}\hat{Y}. \tag{19}$$

Considering

$$\hat{Y} = \mathbf{X}\hat{\beta} \tag{20}$$

and Eq. (13), Eq. (19) can now be expressed as:

$$\hat{Y}^{\star}\hat{Y} = \hat{\beta}^{\star}\mathbf{X}^{\star}Y. \tag{21}$$

Then, the explained sum of squares, *ESS*, (also known as sum squares due to regression), which is defined as the sum of the squares of the differences of the predicted values and the mean value of the dependent (response) variable is:

$$ESS = \hat{\beta}^{\star}\mathbf{X}^{\star}Y - n\left(\bar{Y}^{\star}\bar{Y}\right) \tag{22}$$

where $\bar{Y}$ is the complex number defined by the average values of the real and imaginary parts of the Y vector.

The explained sum of squares is used to calculate the Residual Sum of Squares, *RSS*,

$$RSS = TSS - ESS \tag{23}$$

in which, the total sum of squares, *TSS* is given by

$$TSS = Y^{\star}Y - n\left(\bar{Y}^{\star}\bar{Y}\right). \tag{24}$$

Considering, Eq. (22), the following expression for the *RSS*, is obtained;

$$RSS = Y^{\star}Y - \hat{\beta}^{\star}\mathbf{X}^{\star}Y. \tag{25}$$

The above statistics are useful in quantifying the explanatory power of the model, like the R^2 statistics;

$$R^2 := \frac{ESS}{TSS} = \frac{\hat{\beta}^{\star}\mathbf{X}^{\star}Y - n\left(\bar{Y}^{\star}\bar{Y}\right)}{Y^{\star}Y - n\left(\bar{Y}^{\star}\bar{Y}\right)} \tag{26}$$

or equivalently,

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{Y^{\star}Y - \hat{\beta}^{\star}\mathbf{X}^{\star}Y}{Y^{\star}Y - n\left(\bar{Y}^{\star}\bar{Y}\right)}. \tag{27}$$

As additional statistics, the unbiased estimate, $\hat{\sigma}^2$, of the unknown a posteriori variance of unit weight (variance factor) $\hat{\sigma}^2$ is given by,

$$\hat{\sigma}^2 = \frac{RSS}{n-k} = \frac{\sum \hat{\varepsilon}_i^{\star}\hat{\varepsilon}_i}{n-k}, \tag{28}$$

where $\sum \hat{\varepsilon}_i^{\star}\hat{\varepsilon}_i$ is the *RSS* for the $n \times 1$ observed and $k \times 1$ unknown complex coefficient vectors $\hat{\beta}$ (vector correlates).

4 Testing statistical significance of the estimated vector model parameters

In light of complex number isomorphism leading to the least squares solution of the vector regression model, the following F test (Rencher, 1995 on p. 293) can be used to test the statistical significance of estimated parameters or the predictive power of the vector model.

A partitioned form of the model 1, can be written as:

$$Y = \mathbf{X}\beta + \varepsilon = (\mathbf{X}_1\ \mathbf{X}_2)\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + \varepsilon \tag{29}$$

where $\mathbf{X}$ is a $n{\times}k$ complex matrix, β is a $k{\times}1$ complex vector and, partitioning aims to test the statistical significance of h subset of parameters for which $\mathbf{X}_2$ has h columns and β_2 has h rows. It is assumed that the errors of the observed vectors are homogenous, uncorrelated, Eq. (2), and *normally* distributed (the latter is needed for the implementation of the hypothesis testing).

Under the null-hypothesis for β_2, i.e.

$$\begin{aligned} H_0 &: \quad \beta_2 = 0 \\ H_a &: \quad \beta_2 \neq 0. \end{aligned} \tag{30}$$

Eq. (29) reduces to:

$$Y = \mathbf{X}_1\beta_1 + \varepsilon \tag{31}$$

where $\beta_1 = \left(\mathbf{X}_1{}^*\mathbf{X}_1\right)^{-1}\mathbf{X}_1{}^*Y$. Hence, H_0 can be tested using the following F statistic:

$$F = \frac{(\beta^*\mathbf{X}^*Y - \beta_1{}^*\mathbf{X}_1{}^*)h}{(Y^*Y - \beta^*\mathbf{X}^*Y)(n-k)} = \\ = \frac{([\beta]^T\{\mathbf{X}\}^T[Y] - [\beta_1]^T\{\mathbf{X}_1\}^T[Y])h}{([Y]^T[Y] - [\beta]^T\{\mathbf{X}\}^T[Y])(n-k)} \quad (32)$$

where h and $n-k$ are the corresponding degrees of freedoms. The null-hypothesis is rejected if the calculated F value is larger than the tabulated F value at a given significance level α.

5 Numerical example

Modeling a typhoon track is used here as a numerical example to illustrate the application of the Ordinary Multiple Linear Vector Regression for vector data. The choice of the numerical example is in its intuitive simplicity rather than representing a complex geodetic application. Note that this is an ordinary regression solution as opposed to being a weighted regression solution since the vector data is assumed to be homogeneous (homoscedastic). This is also a multiple regression solution as opposed to simple regression since it involves more than one independent vector variable.

A typhoon is a tropical storm in the region of the Indian or western Pacific oceans. Typhoon paths follow three general directions; straight track, a general westward path affecting the Philippines, southern China, Taiwan, and Vietnam; a parabolic, recurving track, affecting eastern Philippines, China, Taiwan, Korea, and Japan; and the northward track, from point of origin, the storm follows a northerly direction, only affecting small islands (Elsner and Liu, 2003).

The event used in this study is “Vamco” (Figure 1) that occurred in 2009. Vamco appeared at about one thousand kilometers east of Guam on August 16th and vanished at about one thousand kilometers east of Hokkaido on August 25th. It moved northwardly and straightforwardly. The progression of a typhoon can be monitored and expressed as vectors (observed dependent vector variable) in an earth fixed or local reference frame. A number of factors influence the path of a typhoon as independent variables including Sea Surface Pressure Variations (SSPV), which is an independent *scalar* variable, and Geostrophic Wind (GW), which is an independent *vector* variable.

Table 1 lists the two dimensional typhoon displacement vector components (in *km*) in a complex vector Y. In the same table X_1 is the independent scalar variable,

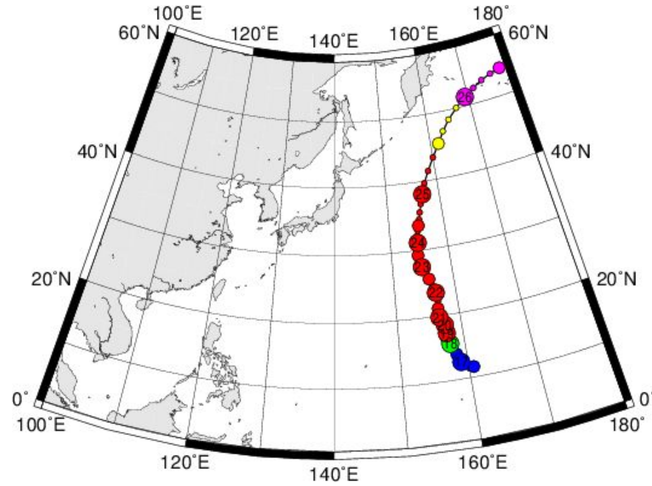

Fig. 1. Vamco’s route (Kitamoto, 2013).

SSPV (in Pascal) and X_2 contains the complex vector components of geostrophic wind (in *km/h*). All data are calculated with 6-hourly intervals using the reanalysis data obtained from NOAA (2012).

5.1 Parameter estimation

Consider now the following vector linear regression model for estimating the 3×1 unknown complex vector regression coefficients β

$$Y = \mathbf{X}\beta + \varepsilon \quad (33)$$

where Y is a 15×1 complex vector, $\mathbf{X}$ is a 15×3 complex design matrix. Using Eq. (13), the estimated vector parameters, $\hat{\beta}$, are readily computed as;

$$\hat{\beta} = \begin{pmatrix} -15.2 + 58.0i \\ -3.3 - 2.9i \\ 20.2 + 12.7i \end{pmatrix}. \quad (34)$$

$\hat{Y}$ can now be calculated using Eq. (20). Table 2 lists the observed complex vector Y, adjusted (predicted) complex vector, $\hat{Y}$, and the complex residual vector $(Y - \hat{Y})$. Using Eq. (28) with $n = 15$ and $k = 3$, the *a posteriori* variance of unit weight is determined to be $\hat{\sigma}^2 = 2993.81$.

The variance/covariance matrix of the vector estimates is obtained from eq. 18. Using the a posteriori variance of unit weight $\hat{\sigma}^2$ the results are:

$$cov\left(\hat{\beta}\right) = \\ = \begin{pmatrix} 1096.76 & -11.46 - 11.57i & 11.49 - 460.29i \\ -11.46 + 11.57i & 0.93 & 6.95 + 11.62i \\ 11.49 + 460.29i & 6.95 - 11.62i & 287.87 \end{pmatrix} \quad (35)$$

Table 1. Sample complex data and their isomorphism.

Track (*km*) Y	Track (*km*) $[Y]$	Const. α	SSPV (*Pascal*) X_1	Geostrophic Wind (*km/h*) X_2	Const. $\{\alpha\}$		SSPV (*Pascal*) $\{X_1\}$		Geostrophic Wind (*km/h*) $\{X_2\}$	
−40.6+133.4i	−40.6	1	−4.7	0.6+2.3i	1	0	−4.7	0	0.6	−2.3
	133.4				0	1	0	−4.7	2.3	0.6
−40.3+100.1i	−40.3	1	−9	−1+1.8i	1	0	−9	0	−1	−1.8
	100.1				0	1	0	−9	1.8	−1
−49.9+100.1i	−49.9	1	−14.7	−0.1+1.3i	1	0	−14.7	0	−0.1	−1.3
	100.1				0	1	0	−14.7	1.3	−0.1
−39.6+111.2i	−39.6	1	−10.2	−0.2+3.1i	1	0	−10.2	0	−0.2	−3.1
	111.2				0	1	0	−10.2	3.1	−0.2
−19.6+111.2i	−19.6	1	−11.2	−0.4+1.9i	1	0	−11.2	0	−0.4	−1.9
	111.2				0	1	0	−11.2	1.9	−0.4
−19.5+89i	−19.5	1	4.8	1.4i	1	0	4.8	0	0	−1.4
	89				0	1	0	4.8	1.4	0
−19.3+77.8i	−19.3	1	−6.3	−0.7+1.9i	1	0	−6.3	0	−0.7	−1.9
	77.8				0	1	0	−6.3	1.9	−0.7
38.3+122.3i	38.3	1	8.1	0.2+1.5i	1	0	8.1	0	0.2	−1.5
	122.3				0	1	0	8.1	1.5	0.2
9.5+122.3i	9.5	1	−10.1	1.4+1.5i	1	0	−10.1	0	1.4	−1.5
	122.3				0	1	0	−10.1	1.5	1.4
37.3+155.7i	37.3	1	−27.8	0.3+2.4i	1	0	−27.8	0	0.3	−2.4
	155.7				0	1	0	−27.8	2.4	0.3
55+211.3i	55	1	−24.3	1.9+2.7i	1	0	−24.3	0	1.9	−2.7
	211.3				0	1	0	−24.3	2.7	1.9
80.1+289.1i	80.1	1	−37	1.1+2.6i	1	0	−37	0	1.1	−2.6
	289.1				0	1	0	−37	2.6	1.1
153.9+366.9i	153.9	1	−54.6	1.3+4.2i	1	0	−54.6	0	1.3	−4.2
	366.9				0	1	0	−54.6	4.2	1.3
242.8+422.5i	242.8	1	−75.8	1.3+4.4i	1	0	−75.8	0	1.3	−4.4
	422.5				0	1	0	−75.8	4.4	1.3
235.7+366.9i	235.7	1	−79.3	2.2+5.2i	1	0	−79.3	0	2.2	−5.2
	366.9				0	1	0	−79.3	5.2	2.2

The R^2 statistics, expressing the explanatory power of the model solution is

$$R^2 = \frac{288689.36}{324615.13} = 0.89 \tag{36}$$

which is obtained from the ratio of eq. 22 for $ESS = 288689.36$ and eq. 24 for $TSS = 324615.13$. This result shows that 89 percent of the variation in typhoon tracks is explained by the SSPV and geostrophic wind.

5.2 Model testing

Using the *F-test* given by eq. 32 in which h represents the degrees of freedom of the estimate (h is equal to 1 in this numerical example), $n-k$ is the degrees of freedom of the errors $(n-k)$ and equal to $15-3=12$, and the isomorphism format of the variables is presented in Table 1 for the second part of the *F test*. The following table can be used for testing the significance of the intercept:

In the above table $\mathbf{X}_1$ is the design matrix without intercept, β_1 is calculated from $\beta_1 = \left(\mathbf{X}_1{}^{\star}\mathbf{X}_1\right)^{-1}\mathbf{X}_1{}^{\star}Y$.

The calculated Fvalue is smaller than the critical value, $F_{(0.05,1,12)} = 4.75$, therefore, the hypothesis of interest, which is the coefficient of intercept equals to zero, cannot be rejected at $\alpha = 0.05$ significance level.

The following table quantifies the test for the significance of the coefficient of SSPV: In the above table $\mathbf{X}_2$ is the design matrix without the intercept, and the sea surface pressure variations, β_2 is calculated from $\beta_2 = \left(\mathbf{X}_2{}^{\star}\mathbf{X}_2\right)^{-1}\mathbf{X}_2{}^{\star}Y$ and, the degrees of freedom of the error term $(n-k)$ is $14-2=12$.

In this case, the calculated F value is larger than the critical value $F_{(0.05,1,12)} = 4.75$, therefore, the hypothesis of interest, which is the coefficient about the SSPV being statistically significant cannot be rejected at $\alpha = 0.05$ significance level.

Table 2. Comparison of observed typhoon track data, calculated typhoon track data, and residuals

		Typhoon Track Data (km)			
Observed		Adjusted (predicted)		Residuals	
Y	$[Y]$	$\hat{Y}$	$[\hat{Y}]$	$\hat{\varepsilon}$	$[\hat{\varepsilon}]$
−40.6+133.4i	−40.6 133.4	−16.5+126i	−16.5 126.0	−24.1+7.4i	−24.1 7.4
−40.3+100.1i	−40.3 100.1	−28.2+108.2i	−28.2 108.2	−12.1−8.1i	−12.1 −8.1
−49.9+100.1i	−49.9 100.1	15.4+126.3i	15.4 126.3	−65.3−26.2i	−65.3 −26.2
−39.6+111.2i	−39.6 111.2	−24.5+148.2i	−24.5 148.2	−15.1−37i	−15.1 −37.0
−19.6+111.2i	−19.6 111.2	−10+124.3i	−10.0 124.3	−9.6−13.1i	−9.6 −13.1
−19.5+89i	−19.5 89	−49+72.2i	−49.0 72.2	29.5+16.8i	29.5 16.8
−19.3+77.8i	−19.3 77.8	−32.4+106.1i	−32.4 106.1	13.1-28.3i	13.1 −28.3
38.3+122.3i	38.3 122.3	−57.2+67i	−57.2 67.0	95.5+55.3i	95.5 55.3
9.5+122.3i	9.5 122.3	27.8+135.8i	27.8 135.8	−18.3−13.5i	−18.3 −13.5
37.3+155.7i	37.3 155.7	53.3+192.2i	53.3 192.2	−16−36.5i	−16.0 −36.5
55+211.3i	55 211.3	70.2+208.3i	70.2 208.3	−15.2+3.1i	−15.2 3.1
80.1+289.1i	80.1 289.1	97.7+233.5i	97.7 233.5	−17.6+55.6i	−17.6 55.6
153.9+366.9i	153.9 366.9	140.3+320.3i	140.3 320.3	13.6+46.6i	13.6 46.6
242.8+422.5i	242.8 422.5	208.5+386.8i	208.5 386.8	34.3+35.7i	34.3 35.7
235.7+366.9i	235.7 366.9	228.3+424.6i	228.3 424.6	7.4-57.7i	7.4 −57.7

Table 3. Numerical values to check the significance of the intercept using F test.

$h = 1$	*SS (reduced)*	$\left(\beta^* X^* Y - \beta_1{}^* X_1{}^* Y\right) h = 9810.21$
$n - k = 12$	*RSS*	$(Y^* Y - \beta^* X^* Y)(n - k) = 2993.81$
F		3.28

Table 4. Numerical values to check the significance of the SSPV coefficient using F test.

$h = 1$	*SS (reduced)*	$\left(\beta^* X^* Y - \beta_2{}^* X_2{}^* Y\right) h = 57042.71$
$n - k = 12$	*RSS*	$(Y^* Y - \beta^* X^* Y)(n - k) = 3811.33$
F		14.97

In the case of testing the statistical significance of the geostrophic wind coefficient, the relevant statistics are as follows:

Table 5. Numerical values to check the statistical significance of the geostrophic coefficient using F test.

$h = 1$	*SS (reduced)*	$\left(\beta^* X^* Y - \beta_3{}^* X_3{}^* Y\right) h = 74078.10$
$n - k = 12$	*RSS*	$(Y^* Y - \beta^* X^* Y)(n - k) = 3811.33$
F		19.44

In the above table $\mathbf{X}_3$ is the design matrix without intercept and geostrophic wind, β_3 is calculated from $\beta_3 = \left(\mathbf{X}_3{}^* \mathbf{X}_3\right)^{-1} \mathbf{X}_3{}^* Y$, and the degree of freedom of the error $(n - k)$ is: $14 - 2 = 12$.

The calculated Fvalue is larger than the critical value, $F_{(0.05,1,35)} = 4.12$. Therefore, the coefficient of geostrophic wind is statistically significant at $\alpha = 0.05$ significance level.

In summary, all the vector model parameters, with the exception of the intercept term, are statistically significant.

6 Conclusion

This study formalized a new approach for the analysis of vector data. The proposed regression of vector data is a new way of thinking for exploratory statistical analysis of various geodetic phenomena such as, polar motion, ocean currents, typhoon/hurricane tracking, crustal deformations, and precursory earthquake signals, and for applications in other disciplines. The observed vector variable of an event (dependent vector variable) is expressed as a function of a number of hypothesized phenomena realized also as vector variables (independent vector variables) and/or scalar variables that are likely to impact the dependent vector variable. Complex numbers are used to represent vector information, and the method of least squares is deployed to estimate the vector model parameters after transforming the complex vector regression model into a real vector regression model through isomorphism and back again to a complex vector formulation solution. In addition, various operational statistics for testing the predictive significance of the estimated vector parameter coefficients are derived and the solution process is demonstrated using a simple numerical example.

The proposed vector regression modeling, the derivation of its solution, and the relevant statistics elucidated in this study are by no means exhaustive. For example, the exploratory nature of the vector regression can be easily extended to time series analysis and forecasting. Furthermore, a number of assumptions were made to simplify the vector model development, such as homogeneity of the vector errors and absence of correlation among them. Given the fact that the vector quantities are usually derived from location information such an assumption can be violated and additional model development will be needed for optimal solutions in future studies.

The assumption of “no error in variables” i.e. in the presence of stochastic vector variables, may be violated for a number of application of the vector regression models. A solution approach similar to classical total least squares approach can be developed to handle such problems.

Acknowledgement: We thank the editors of the Journal of Geodetic Science who were supportive and open minded, about the potential usefulness of the formulation and the solution methodology introduced in this study, and the two anonymous reviewers for their time and effort.

References

Anderson H. H., Hojbjerre M., Sorensen D. and Erikson P. S., 1995, Linear and Graphical Models, Springer-Verlag, New York.

Chinese National Meteorology Department – Shanghai Typhoon Research Institution Centre, 2012, China Typhoon Website. Available from: China, Chinese national meteorology department Web site: http://www.typhoon.gov.cn/index.php?controller=spage{&}pid=169[Accessed: May 14, 2012].

Elsner J.B. and Liu K.B., 2003, Examining the ENSO–typhoon hypothesis, CLIMATE RESEARCH, Vol. 25, pp. 43-54.

Kitamoto A., 2013, http://agora.ex.nii.ac.jp/~kitamoto, Research Organization of Information and Systems (ROIS) / National Institute of Informatics (NII) Department: Digital Content and Media Sciences Research Division.

NOAA, 2012, NOAA Earth System Research Laboratory. Available from: USA, NOAA Web site: http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html [Accessed: May 14, 2012].

Rencher A.C., 1995, Multivariate Statistical Inference and Applications, John Wiley& Sons, New York.

Schneider D., 1982, Complex crustal strain approximation, Dept. of Surveying, University of New Brunswick, Fredericton.

Su B.K., 1988, Statistical Diagnosis and Forecast in Atmospheric Science, 1sted, Nanjing University, Jiang Su. ISBN: 7-305-00039-6.

Tobler W., 1994, Bidimensional regression. Geographical Analysis, 26, 187–212.

Wilson E.B., 1901, Vector Analysis, Yale University Press, pgs. 436.