EAI Endorsed Transactions
on Context-aware Systems and Applications
Research Article EAI.EU

Kriging interpolation model: The problem of predicting the number of deaths due to COVID-19 over time in Vietnam

Nguyen Cong Nhut

Faculty of Information Technology at Nguyen Tat Thanh University
300A Nguyen Tat Thanh street, Ward 13, District 4, Ho Chi Minh City, Vietnam

Abstract

The COVID-19 pandemic can be considered a human disaster, it has claimed the lives of many people. We only know the number of deaths due to COVID-19 through government statistics, but on days when there are no statistics, how do we know whether people died that day or not? This study aims to predict the number of new deaths per day due to COVID-19 in Vietnam on days when observational data is not available and predict the number of deaths in the future. The study used COVID-19 data from the World Health Organization (WHO). A total of 260 days were collected and the author processed and standardized the data. Based on available data, the author uses Kriging interpolation statistical method to build a forecast model. As a result, the author has selected a prediction model suitable for a highly reliable data set, the regression coefficient and correlation coefficient are close to 1, the error between the model's prediction results compared to data. There are days when the prediction error is almost zero. The study has built a future forecast map of the number of new deaths per day due to COVID-19. The article concludes that applying the Kriging statistical method is appropriate for COVID-19 data. This research opens up new research directions for related fields such as earthquakes, mining, groundwater, environment, etc.

Received on 15 September 2023; accepted on 24 September 2023; published on 25 September 2023

Keywords: Geostatistics, COVID-19, Kriging, Statistical, Interpolation

doi:10.4108/eetcasa.v9i1.3954

1. Introduction

The COVID-19 pandemic is a matter of great concern around the world. The COVID-19 pandemic is an infectious disease pandemic caused by the SARS-CoV-2 virus and its variants that is taking place on a global scale. Originating in late December 2019 with the first outbreak in Wuhan city in central China, COVID-19 originated in a group of people with pneumonia of unknown cause. On March 11, 2020, the World Health Organization (WHO) issued a statement calling "COVID-19" a "Global Pandemic".

Governments around the world have responded to protect the health of people and community groups around the world, including: restricting movement, blocking quarantine, declaring a state of emergency, using curfews, implementing social distancing, canceling mass events, closing schools and less important business and service establishments, encouraging people to raise their own awareness of prevention, wear a mask, limit going out when not necessary, and at the same time transform the business, study and work model from traditional to online. The worldwide effects of the current COVID-19 pandemic include: loss of life, economic and social instability.

In the world, there have been many studies related to the topic of COVID-19. Topics are often focused around such issues as transmission origin research [12]. Applying the Kriging statistical method to time forecasting has also been studied in some studies, such as: Using geostatistics to analyze spatial and temporal Variations of groundwater levels [1], [6]. Analysis of the Spatio-Temporal variation of groundwater levels using geostatistics [7]. Using CoKriging method to predict air pollution [10], [11]. Application of Geostatistics to forecast DO pollutants in water [9]. Analyzing incomplete spatial data in air pollution prediction [8].

*Corresponding author. Email: ncnhut@ntt.edu.vn

In Vietnam, in response to the outbreak of the COVID-19 pandemic, the state has taken measures to control the epidemic, such as limiting mass gatherings, restricting travel, social distancing, wearing masks, disinfection and vaccination. Until now, there have been no domestic or international studies on the problem of predicting This study aims to predict the number of new deaths per day due to COVID-19, the latest research only stops at the problem of estimating the number of infections. Therefore, in this study, the author introduces the Kriging statistical method and its application in predicting the number of new deaths per day due to COVID-19.

The research objectives are:

1. Build models based on measured data.
2. Forecasting This study aims to predict the number of new deaths per day due to COVID-19 without statistics.
3. This study uses the Kriging statistical method to forecast This study aims to predict the number of new deaths per day due to COVID-19 in Vietnam.

2. Literature Review

Kriging

In geostatistics, Kriging, also known as Gaussian process regression, is a method of interpolation based on Gaussian processes governed by prior covariances. Under suitable assumptions of the prior, kriging gives the best linear unbiased prediction (BLUP) at unsampled locations [3].

Interpolating methods based on other criteria such as smoothness (e.g., smoothing spline) may not yield the BLUP. The method is widely used in the domains of spatial analysis and computer experiments.

The theoretical basis for the method was developed by the French mathematician Georges Matheron in 1960, based on the master's thesis of Danie G. Krige, the pioneering plotter of distance-weighted average gold grades at the Witwatersrand reef complex in South Africa. Krige sought to estimate the most likely distribution of gold based on samples from a few boreholes.

In general, previous studies mainly focused on building predictive models in terms of space using the Co-Kriging method, but no studies have built predictive models in time series.

Gaussian process

A Gaussian process is a stochastic process such that every finite set of such random variables has a multivariable normal distribution. The distribution of the Gau process is the point distribution of all the random variables [2].

In this study, I applied the Kriging method as a management and decision support system tool to analyze the time variations of the COVID-19 epidemic so that I could gain a better picture of the COVID-19 death toll over a long period of time.

Research Questions:

With the aim of understanding how the number of deaths due to the COVID-19 pandemic changed over time, the study sought to answer the following questions:

1. Do the days without statistics predict the number of deaths?
2. Over time, will the number of deaths from COVID-19 be high in the future?
3. Is the forecasting model reliable enough?
4. Is the COVID-19 epidemic in Vietnam really under control?

3. Methods

Data Collection and Analysis

The overall objective of this study was Covid-19 data from the World Health Organization (WHO), with a total of 880 days collected (July 31, 2020 to December 29, 2022). The convenience sampling method was used to select a 95% confidence interval with an error of 5%, and after removing invalid values, we got a sample size of 279 values with continuous collection time in days from July 26, 2021 to April 30, 2022 Table 1. Of the 259 days of observed data that the author included in the model building, the remaining 20 days were not available, but data were not available (NA). An overview of the COVID-19 death toll data is depicted in Table 2.

The author proceeds to process and normalize Figure 1 data by histogram. In Figure 1, the left histogram before normalizing the data is skewed to the left, and the histogram on the right of the data has been more balanced and normalized.

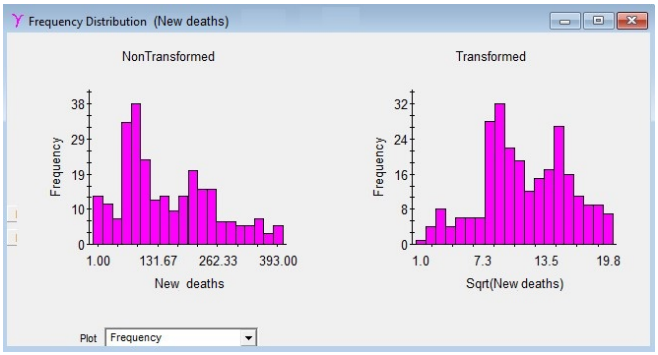

Figure 1. Standardization of COVID-19 death toll data

Table 1. Data on COVID-19 deaths by day

Date	New deaths	Date	New deaths
7/26/2021	154	...	...
7/27/2021	NA	4/14/2022	23
7/28/2021	106	4/15/2022	23
7/29/2021	NA	4/16/2022	10
7/30/2021	139	4/17/2022	10
7/31/2021	145	4/18/2022	13
8/1/2021	NA	4/10/2022	18
8/2/2021	NA	4/20/2022	7
8/3/2021	190	4/21/2022	9
8/4/2021	256	4/22/2022	7
8/5/2021	393	4/23/2022	6
8/6/2021	296	4/24/2022	9
8/7/2021	234	4/25/2022	8
8/8/2021	NA	4/26/2022	8
8/9/2021	360	4/27/2022	5
8/10/2021	388	4/28/2022	3
8/11/2021	342	4/29/2022	1
8/12/2021	326	4/30/2022	3
8/13/2021	275	5/1/2022	NA
8/14/2021	349	5/2/2022	NA
8/15/2021	337	5/3/2022	NA
8/16/2021	367	...	...

Table 2. Overview of COVID-19 death toll data

Date	New_deaths
Length: 279	Min. : 1.0
Class :character	1st Qu.: 74.0
Mode :character	Median :123.0
	Mean :150.4
	3rd Qu.:222.0
	Max. :393.0
	NA's :20

Figure 2 shows the distribution of the data over time, the horizontal and vertical axes represent time in days, and the transverse axis shows the number of deaths from COVID-19. Different symbols and colours represent different death tolls.

Research Methodology

In geostatistics, the main tool is the variogram, which represents the spatial dependence between observations. The variogram $(2\gamma(h))$ is defined as an expectation of the random variable $[Z(u) - Z(u+h)]^2$

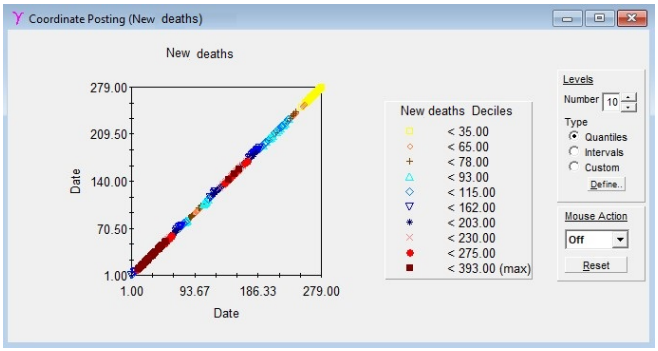

Figure 2. Distribution of COVID-19 death toll data over time

[1], that is:

$$2\gamma(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} [Z(u_i) - Z(u_i + h)]^2, \qquad (1)$$

where h is the distance between points in space. $Z(u_i), Z(u_i + h)$ are observed values at position u_i and $u_i + h$. $N(h)$ is the number of pairs of points.

According to the quadratic stationary conditions [13], [4] one obtains

$$E(Z(u)) = \mu,$$

and the covariance

$$Cov[Z(u), Z(u+h)] = E[Z(u)Z(u+h) - \mu^2] = C(h). \quad (2)$$

Then $2\gamma(h) = E[Z(u) - Z(u+h)]^2 = C(0) - C(h)$.

In geostatistics, there are four most commonly used models: Linear, Spherical, Exponential and Gaussian. Based on an experimental variogram, a variogram model that fits the data is selected using a technique called cross - validation.

The cross plot between the estimated value and the actual value shows the correlation coefficient r^2. The best fit variogram model is selected based on the highest correlation coefficient and is approximately equal to 1.

The Kriging method is a group of geostatistical methods used to interpolate the data of a random field at an unknown point from known values at neighboring points [13]. In the Kriging method, there are two types of Simple Kriging and Ordinary Kriging. The Simple Kriging method is the Kriging method for which the mean μ is known in advance, the formula is as follows

$$\widehat{Z}(u_0) = \sum_{i=1}^{n} \lambda_i Z(u_i) + \left[1 - \sum_{i=1}^{n} \lambda_i\right] \mu,$$

The Ordinary Kriging method is the Kriging method of unknown mean, based on the hypothesis of a truly stable stochastic function.

Table 3. Isotropic variogram values of New deaths

	Nugget	Sill	Range	r^2	RSS
Linear	6.988	21.149	190.118	0.508	158
Gaussian	**0.04**	**18.01**	**79.3279**	**0.95**	**26.2**
Spherical	0.01	17.95	100.7	0.945	34.8
Exponential	0.01	18.8	135	0.867	81.4

$$\hat{Z}(u_0) = \sum_{i=1} \lambda_i Z(u_i). \quad (3)$$

Kriging minimizes the mean squared error

$$\min \sigma_e^2 = [Z(u_0) - \hat{Z}(u_0)]^2.$$

For second order stationary process the last equation can be written as

$$\sigma_e^2 = C(0) - 2\sum_{i=1} \lambda_i C(u_0, u_i) + \sum_{i=1}\sum_{j=1} \lambda_i \lambda_j C(u_i, u_j) \quad (4)$$

subject to $\sum_{i=1} \lambda_i = 1$

Therefore the minimization problem can be written

$$\min \left\{ C(0) - 2\sum_{i=1} \lambda_i C(u_0, u_i) + \sum_{i=1}\sum_{j=1} \lambda_i \lambda_j C(u_i, u_j) - 2\beta(\sum_{i=1} \lambda_i - 1) \right\} \quad (5)$$

where β is the Lagrange multiplier. After differentiating (5) with respect to $\lambda_1, \lambda_2, ..., \lambda_n$, and β and set the derivatives equal to zero we find that

$$\sum_{j=1} \lambda_j C(u_i, u_j) - C(u_0, u_i) - \beta = 0, \; i, j = 1, 2, ..., n$$

and $\sum_{i=1} \lambda_i = 1$ [8], [9], [10], [11].

4. Results

To check the anisotropy of COVID-19, the author compares the histogram of variation in many directions [5]. In this study, four main directions were used, namely $0^0, 45^0, 90^0$, and 135^0 with an angular tolerance of $\pm 45^0$ used to determine anisotropy.

Figure 3 shows that of the 4 models, the model curve Gaussian is in good agreement with the experimental data. From Figure 3, we have the values in the Gaussian model [$Nugget = 0.04; Sill = 18.01; Range = 79.3279; r^2 = 0.95$]. It shows the best-fit omnidirectional variogram in deaths over time obtained based on cross - validation. Based on the variogram map of the new death parameters, the isotropic variogram model is suitable. The values for the four models are presented in Table 3.

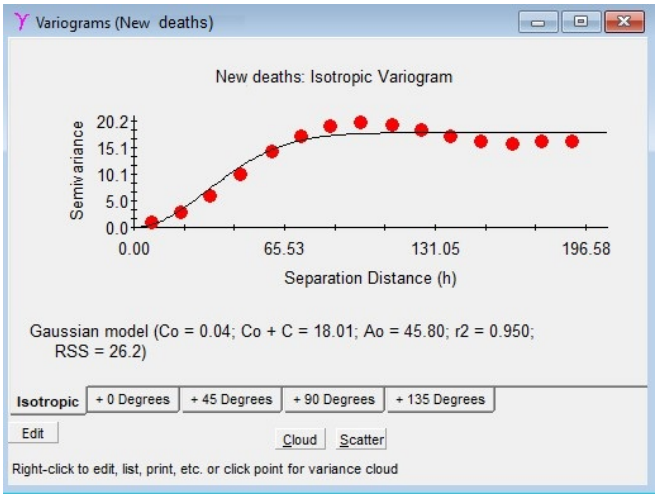

Figure 3. Fitted variogram for the temporal analysis of New deaths

Table 4. Model test results

Coefficient of regression	Coefficient of correlation	Standard Error	Standard Error Prediction
0.988	0.933	0.017	25.767

Table 3 is the result of 4 models. To choose which model is the best, it is based on two important criteria: RSS (Residual Sums of Squares) and r^2 (coefficient of determination). The RSS provides an accurate measure of how well the model fits the variance data; The lower of RSS, the more suitable the model, among the 4 models, the Gaussian model has the smallest RSS = 26.2. The r^2 provides an indicator of how well the model fits the variance data. However, r^2 is not as strong as the RSS value [8], [9], [10], [11].

Model Testing

Criteria for evaluating whether the selected model fits the data is based on regression coefficient and correlation coefficient. The model test results are shown in Table 4 and Figure 4. The regression coefficient and the correlation coefficient are close to 1, the standard error is approximately zero. The comparison results between the estimated value and the actual value, shown in Figure 4. Conclusion that the selected model is suitable, the error is small, scatter plot of the New deaths parameter in Figure 5.

Figures 6 and 7 are Kriging interpolation maps of COVID-19 deaths. The highest number of deaths is in white, and decreasing to the lowest is in blue. In the same color, the number of deaths is nearly equal, with only a small difference between them.

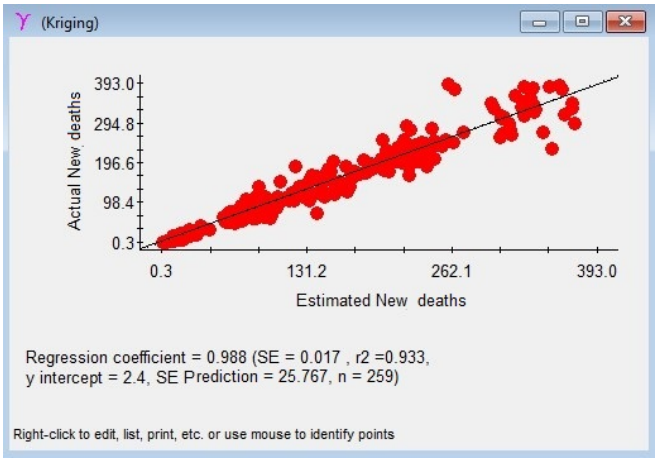

Figure 4. Test results fail to predict new deaths

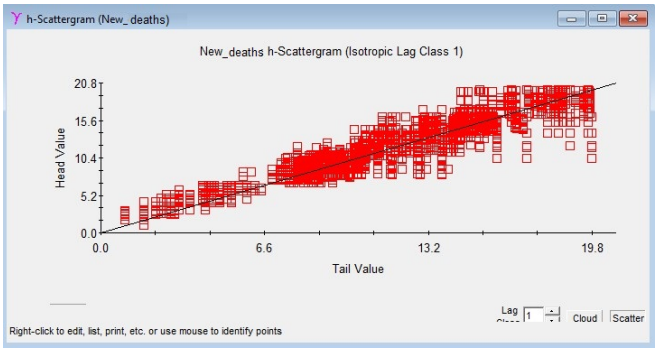

Figure 5. Scatter plot of the New deaths parameter

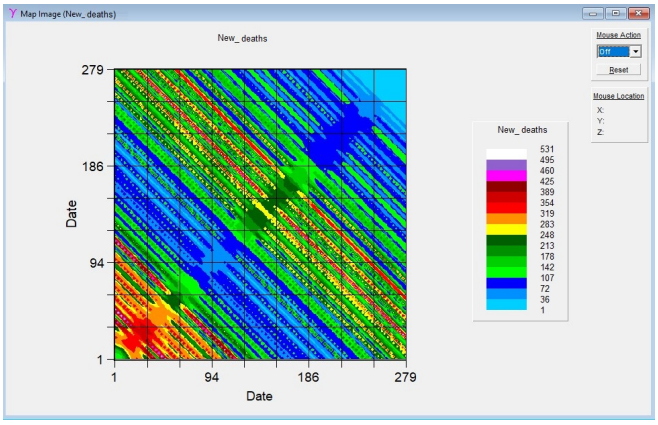

Figure 6. 2D Kriging interpolation map of the number of new deaths per day due to COVID-19

5. Discussion

Based on the maps in Figures 6 and 7, we see that the further we get to the last milestone of the data, 279 (April 30, 2022), the lowest number of deaths, there are almost no deaths due to COVID-19 in Vietnam. This is also completely true of the fact that, after April 30, 2020, the COVID-19 situation in Vietnam will be

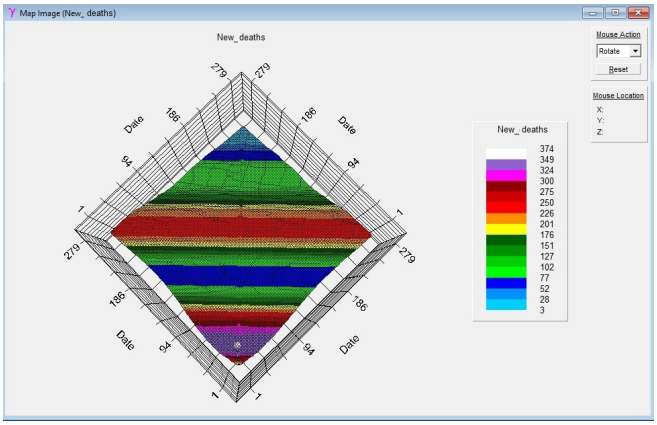

Figure 7. 3D Kriging interpolation map of the number of new deaths per day due to COVID-19

Table 5. Forecast of COVID-19 deaths in Vietnam

Date Date	Actual Actual	Estimated (Kriging)	Estimated (IDW)	Error (Kriging)	Error (IDW)
...	...	...	...	...	...
4/2/2022	37	37.32	34.71	0.32	2.29
5/2/2022	37	36.74	34.47	0.26	2.53
6/2/2022	42	34.48	33.03	7.52	8.97
8/2/2022	39	34.18	30.26	4.82	8.74
9/2/2022	31	32.37	29.07	1.37	1.93
10/2/2022	21	31.03	29.57	10.03	8.57
11/2/2022	35	24.17	25.21	10.83	9.79
12/2/2022	26	23.82	25.46	2.18	0.54
13/2/2022	19	24.65	24.28	5.65	5.28
14/2/2022	17	24.56	23.58	7.56	6.58
15/2/2022	28	20.27	20.24	7.73	7.76
...	...	...	...	...	...

strictly controlled. Due to drastic measures taken by the government to control the disease and people's awareness of disease prevention, the death toll from COVID-19 has almost disappeared.

Compare the Kriging method with other traditional methods such as the sample averaging and distance inverse methods. The sample mean is the average value of sample values close to the location to be estimated. Inverse distance weighting (IDW) is a deterministic method for multivariate interpolation with a known set of scatter points. The values assigned to the unknown points are calculated as a weighted average of the available values at the known points. Table 5 shows the forecast results for the Kriging method with traditional methods. The forecast error between the observed value and the actual value by the Kriging method is smaller than the error when using the IDW method.

The model's forecast results also have errors, which may be due to inaccurate statistics on the number of new deaths per day due to COVID-19 in some days

or deaths due to underlying diseases. Therefore, in the next study, the author will study the effect of the underlying disease on the mortality rate due to COVID-19 to reduce the error in prediction.

6. Conclusion

With the data set that the author is studying, the Gaussian model is suitable, and the indicators are better than the remaining models. The error between the estimated values and the actual value of the model is very small (0.017). The regression coefficient is equal to 0.988, and the correlation coefficient is 0.933 (approximately 1.0), showing that the choice of interpolation model is appropriate.

Using the Kriging forecasting method to predict COVID-19 deaths for days with no observed data or days with missing data (NA), predicting future COVID-19 deaths results in a very small error between the estimated value and the actual value. The study shows that the efficiency, reasonableness, and high reliability of the Kriging method to build a predictive model are appropriate. When building models, attention should be paid to model error values, object-specific data, and the results of model selection to choose a suitable model for actual data from models providing different accuracy. Therefore, experience in model selection plays an important role in research results.

Acknowledgment

It is possible that the author's research capacity is limited, so there will be some shortcomings. Therefore, the author wishes to receive sincere suggestions from scientists to improve and produce better results.

References

[1] Ahmadi, S.H. and Sedghamiz, A. Geostatistical analysis of spatial and temporal variations of groundwater level. *Environmental Monitoring and Assessment*, 129(1-3), 2007.

[2] Bishop, C.M. Pattern recognition and machine learning. *Springer*, 2006.

[3] Chung, S.Y., Venkatramanan, S., Elzain, H.E., Selvam, S., and Prasanna, M. V. Supplement of missing data in groundwater-level variations of peak type using geostatistical methods. *GIS and Geostatistical Techniques for Groundwater Science*, 2019.

[4] Gentile1, M., Courbin, F., and Meylan, G. Interpolating point spread function anisotropy. *Astronomy and Astrophysics manuscript no. psf˙interpolation*, 10 2012.

[5] Goovaerts, P. Geostatistics for natural resources evaluation. *New York: Oxford University Press*, 1997.

[6] Hua, Z., Cheng, W., Yi, S., and Qing, J. Geostatistical analysis of spatial and temporal variations of groundwater depth in shule river. *Wase International Conference on Information Engineering, China*, 8 2009.

[7] Mini, P.K., Singh, D.K., and Sarang, A. Spatio-temporal variability analysis of groundwater level in coastal aquifers using geostatistics. *International Journal of Environmental Research and Development*, 4(4), 2014.

[8] Nhut, N.C, and Man, N.V.M. Analyzing incomplete spatial data in air pollution prediction. *Journal Southeast-Asian J. of Sciences*, 6(2), 2018.

[9] Nhut, N.C. Applying geostatistics to predict dissolvent oxygen (do) in water on the rivers in ho chi minh city. *The 8th International Conference on Context-Aware Systems and Applications, and Nature of Computation and ommunication, ICCASA 2019, My Tho, Vietnam*, 2019.

[10] Nhut, N. C., Man, N.V.M., and Phu, V.L. Co-kriging method for air pollution prediction: A case study in saigon. *Thailand Statistician*, 2020.

[11] Nhut, N.C. Applying cokriging method for air pollution prediction pm10 in binh duong province. *Context-Aware Systems and Applications, and Nature of Computation and Communication*, 2021.

[12] Page, J. Virus sparks soul-searching over china's wild animal trade. *Wall Street Journal*, 2020.

[13] Webster, R., and Oliver, M.A. Geostatistics for enviromental scientists. *2nd Edition, John Wiley and Sonc LTD, The Atrium, Southern Gate, Chichester, West Sussex PO19, England*, 2007.