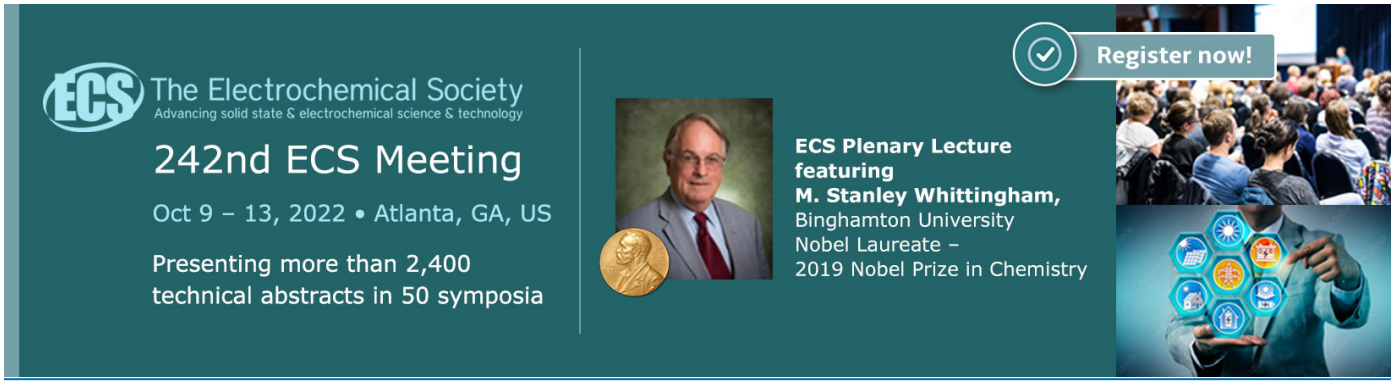
The prevalence of Helicobacter pylori in referral population of Turkey

To cite this article: Komiljon Usarov *et al* 2021 *J. Phys.: Conf. Ser.* **1988** 012092



Journal of Physics: Conference Series **1988** (2021) 012092 doi:10.1088/1742-6596/1988/1/012092

The prevalence of Helicobacter pylori in referral population of Turkey

Komiljon Usarov[1], Anvarjon Ahmedov[2], Mustafa Fatih Abasiyanik[3], Ku Muhammad Na'im Ku Khalif[4] and Abdulkasim Akhmedov[5]

[1,2,4,5]Centre for Mathematical Sciences, College of Computing & Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Pahang, Malaysia
[3]Pritzker School of Molecular Engineering, The University of Chicago, Edward H. Levi Hall, 5801 South Ellis Avenue, Chicago, Illinois 60637, USA

[1]**Email:** ukomiljon@gmail.com

Abstract. Helicobacter pylori infection is commonly associated with gastroduodenal diseases in humans, such as chronic gastritis and peptic ulcers, gastric mucosa-associated lymphoid tissue lymphoma, and even gastric cancer, which leads to high cost to society for treatment and even to death many people, when people do not know early of the infection prevalence. In this work we proposed a forecasting model to predict the infection prevalence. Based on our results society can make simple early prevention acts against the infection. The early prevention acts decrease the cost of treatment and save many people's lives in the world.

Keywords: *H. pylori*, infectious disease prediction, multivariate linear regression

1. Introduction

H. pylori causes persistent infection which develops inflammation and leads to peptic ulcer, chronic gastritis, gastric mucosa and gastric cancer in the human stomach [1],[2],[3]. According to GLOBOCAN 2018 data, stomach cancer is the 3rd most deadly cancer which estimated 783,000 deaths in 2018 [4]. Moreover, the infection is highly prevalent in approximately 50% of the population in the world [1]. It causes high cost to society and even brings on high risk of humans' lives.

To prevent prevalence of the infection, it is crucial to know early about the infection prevalence using forecasting models. Literature consists of some interesting research work related to forecasting models for infections such as Malaria, Scarlet fever, Chickenpox [5] combining Big Data and Neural Network. Moreover, there are some articles about predictions based on environmental factors which have a great impact on the prevalence of the infections. For instance, [6] built a time series model based on eight climate variables to predict hand, foot and mouth disease. In addition, [1] showed that average daily sunshine time correlated positively with H. pylori infection [1]. Based on the previous studies we conclude that climate variables can be used to obtain more accurate and efficient prediction of infections prevalence.

However, there is no prediction model created for H. pylori infection prevalence in the existing literature which is considered with climate variables yet. Therefore, the main purpose of this study is to design a forecasting model for H. pylori (Hp) infection prevalence based on Humidity (H), Dew Point (DP), Temperature (T), Wind Speed (W) and Pressure (P). The forecasting model is used with multivariate linear regression model (MVLR).

We have derived a prediction of the H. pylori infection prevalence by using forecast modelling. Due to the results of our prediction infectious doctor can produce recommendations on early information about the spreading of the infection. It gives a chance to act for prevention procedures against the infection, which leads not only to reduce the prevalence of the infection, but it also minimizes social costs for the public and saves many people's lives.

2. Materials and methods

2.1. Research Data

Data were extracted from Microsoft Excel which were recorded from 1999 to 2003 in the Samatya hospital in Istanbul. The data included 48 attributes of patients such as date of visitors (date of date type), age (numeric of date type), sex*, smoking*, alcohol*, duodenal ulcer (DU*), esophagus ulcer (EU*), gastrıc ulcer (GU*), peptıc ulcer (PU*), gastrit*, pain*, stomach ache*, abdominal pain*, vomit*, aspirin usage*, gastrit cancer* and CLO* with 4388 patients.

(* Binary data type of attributes where there is only 0 or 1 value). Weather data (WD), including H(%), DP(°F), T(°F), P(Hg), W(mph) was obtained from historical data by average daily information in the https://www.wunderground.com/history website, joined with the visitor date attribute. The joined data transformed to weekly (253 rows) and monthly (53 rows) data by aggregations COUNT of CLO, SUM of CLO (Hp), m of H, DP, T, P, W attributes where were achieved total CLO infected patients (Hp), total of patients, m of H, DP, T, P, W, respectively. The final data consists of the total number of patients, total number of CLO infected patients (Hp), H (m of Humidity), DP (m of dew point), T (m of temperature), P (m of Pressure), W (m of wind speed) for the weekly and monthly dataset which visualize in table 1.

Table1. Monthly transformed data

	Hp	H	DP	T	P	W
1	22	78.05	65.26	73.18	29.52	10.18
2	15	79.57	72.59	80.01	29.59	11.30
3	19	75.22	69.46	78.52	29.76	10.94
4	62	67.54	58.76	70.73	29.83	8.15
5	60	77.42	56.43	63.79	29.91	8.77
6	34	74.36	45.57	53.86	29.97	12.44
7	82	77.33	43.99	51.50	29.80	12.84
8	51	69.07	21.92	35.61	29.81	10.01
9	40	71.38	32.28	45.47	29.96	10.03
10	21	60.88	44.78	60.53	29.89	5.30
11	67	71.14	49.17	59.37	29.65	7.22
12	68	59.06	44.30	60.25	29.93	11.16
13	39	57.38	52.77	70.04	29.95	10.04
14	28	55.64	60.22	78.71	29.77	9.59
15	24	70.41	66.10	76.98	29.86	13.36
16	51	68.48	58.71	70.52	29.82	9.35
17	37	77.51	49.77	57.19	30.03	8.74
18	73	75.86	49.12	57.21	29.98	7.10
19	31	77.76	44.46	51.46	30.06	6.91
20	53	72.04	36.83	46.87	30.11	10.51
21	49	68.86	34.18	47.30	29.91	10.08
22	24	68.53	45.81	56.62	29.66	11.35
23	54	70.14	46.74	57.44	29.83	9.19
24	96	64.67	49.13	62.26	29.87	8.80
25	54	53.86	52.92	72.09	29.76	10.21
26	61	63.31	64.74	79.58	29.77	11.21
27	27	74.90	68.13	77.12	29.75	6.91
28	53	68.82	59.93	71.78	29.83	9.99
29	26	77.53	60.67	68.49	30.05	10.46
30	1	73.20	50.00	59.40	30.10	3.50
31	8	91.47	37.98	41.71	29.76	13.37
32	44	81.58	29.86	36.61	30.16	7.89
33	30	80.49	39.96	46.85	30.03	8.07
34	57	79.59	43.00	49.95	29.81	9.26
35	25	79.55	46.29	53.41	29.71	8.74
36	42	72.85	53.69	63.68	29.85	10.81
37	105	68.53	61.36	73.48	29.79	10.82
38	113	68.50	69.00	81.66	29.68	8.16
39	54	71.97	64.75	75.21	29.78	11.57
40	51	77.16	62.49	70.76	29.81	7.13
41	47	80.62	54.82	61.19	29.83	8.82
42	50	81.82	49.61	55.53	30.03	6.17
43	50	68.14	23.86	37.73	30.20	10.81
44	73	87.82	42.42	46.77	29.75	11.96
45	37	90.55	25.00	35.92	29.84	17.27
46	2	85.86	33.70	40.88	29.94	12.36
47	0	73.50	39.95	48.75	29.55	9.90
48	16	69.90	54.07	65.11	29.82	9.04
49	74	57.26	57.69	74.82	29.82	10.30
50	60	59.47	62.04	77.97	29.82	9.76
51	30	56.33	61.51	79.53	29.80	11.50
52	74	64.82	55.81	68.60	29.88	10.47
53	15	67.44	56.52	68.45	29.70	9.91

2.2. Method

The forecast of H. pylori infection prevalence is modelled by multivariate linear regression:

$$y = g(H, D, T) =$$

$$\beta_1 sin^2(\rho_1 H + \rho_2) sin^2(\rho_3 T + \rho_4) + \beta_2 sin^2(\rho_5 H + \rho_6) sin^2(\rho_7 T + \rho_8) +$$

$$\beta_3 sin^2(\rho_9 D + \rho_{10}) sin^2(\rho_{11} W + \rho_{12}) + \beta_4 sin^2(\rho_{13} D + \rho_{14}) sin^2(\rho_{15} W + \rho_{16}) + \beta_5 \quad (1)$$

where:

y - dependent variable (Hp – the number patients who had positive CLO infection test),
H - the average of humidity (%)
D - the average of dew point (°F)
T - the average of temperature (°F)
ρ_{1-16} - non-linear regression coefficients,
β_{1-4} - regression coefficients,
β_5 - constant.

By minimizing the residual sum of squares, we determined the best fit of ρ_{1-16} and β_{1-5} vector paraments to the data in table 1. The residuals are defined for each observed data-point as

$$\varepsilon_i = y_i - g(H_i, D_i, T_i) \quad (2)$$

where y_i is the number of the total H. pylori infected per month. We used nonlinear least square solver of SciPy in python from scipy.org.

2.3. Model assumptions

To achieve validness of the tests of hypothesis (like t-test and F-test) and to enhance that ordinary least squares (OLS) estimators are the Best Linear Unbiased Estimator (BLUE), it is required to follow four base assumptions:

1. The relationship between the dependent variable and the independent variables is linear.

2. The residuals are independent.
   The Durbin-Watson statistic T_{DW} was used to check that residuals are independent.
   If T_{DW} is between 1.65 and 2.35, there is no autocorrelation.
   If T_{DW} is between 1.21 and 1.65 or between 2.35 and 2.79, the test is inconclusive [7]

3. Homoscedasticity.
   Homoscedasticity is a word used for the “constant variance” assumption. The regression model assumes that the residuals have the same variance throughout. When this assumption is violated, the problem is called “heteroscedasticity,” or changing variance. We used the Breusch – Pegan and White test to check heteroscedasticity.

4. Normality of residuals with mean equals to zero.
   Errors need to be a normal probability distribution. This makes no difference to the estimates of the coefficients, or the ability of the model to forecast. But it does affect the F-, t-tests and confidence intervals.

3. Results & Discussions

3.1. Model

The following formula is obtained by applying optimal parameters of the proposed model (1), which is described for Hp infection prevalence:

$$y = f(H, D, T) = 84.3107sin^2(1.4246\,H + 55.2660)sin^2(3.9785\,T{+}7.7484)$$

$$-58.0112sin^2(7.1189\,H - 0.2237)sin^2(8.7530\,T{+}26.8308)$$

$$+73.5578sin^2(11.6994\,D{-}26.9552)sin^2(13.9345\,T{+}20.1594)$$

$$+39.4570sin^2(16.9565\,D{+}16.4408)sin^2(19.0438\,T{+}18.1347) + 14.5355 \quad (3)$$

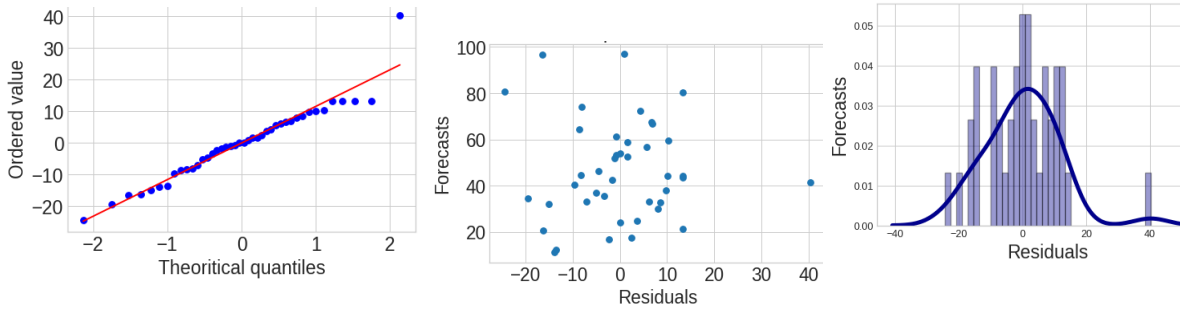
We obtained a MVRM to predict prevalence of the infection based on climate variables, where H. pylori is dependent variable and H, DP and T are independent variables. This formula predicts the accuracy with coefficient of determination (COD) equals to 0.77 and 0.75 for train data (42 months) and test data (11 months), respectively.

With training of subset of data we obtained a Multi Varies Regression Model which allow us to predict the accuracy with coefficient of determination (R2) equals 77% and 75% for train data (42 months) and test data (11 months), respectively (figure 4). The adjustment R2 is 75% which is comparatively very high, means that the correlation coefficient between the observed value of the dependent variable and the forecast value based on the regression model is very strong.

The value of F statistic is 31.39 and the significance of F is 0, endorsed by ANOVA, which is less than the critical value ($p<0.001$). The significance of the model is verified by rejection of the null hypothesis.

3.2. Model assumptions

The first assumption is consequence of the linearity of the β_{1-5} coefficients in the proposed model. From DW test result 1.717 we conclude that there is no autocorrelation between residuals and predicted values (figure 1.b.). The third assumption is verified by Breusch–Pegan Test, the null hypothesis was not rejected ($p>0.05$). It can be seen in figure 1.a. by the QQ plot, which easily proves that it is homoscedasticity. The last assumption is also true for the given model and the mean of residuals is zero. In addition, Shapiro-Wilk, Anderson-Darling tests show that null hypothesis is rejected ($p>0.05$), which means the residuals are normally distributed (figure.1.c).



a. QQ Plot. Heteroscedasticity b. Independence of residuals c. Distribution of residuals

Figure 1. Residuals analysis

3.3. Prediction Results.

The forecasts result of training and testing data were represented by figure 2, where it was separated by a grey vertical line. By date (month, year) and the number of CLO are represented by x-axis and y-axis, respectively. Actual data is in blue colour, training data is in green and testing data is in red colour. The forecasting data started from November 2002 till October 2003 which means that almost one year forecasts is high accurate.

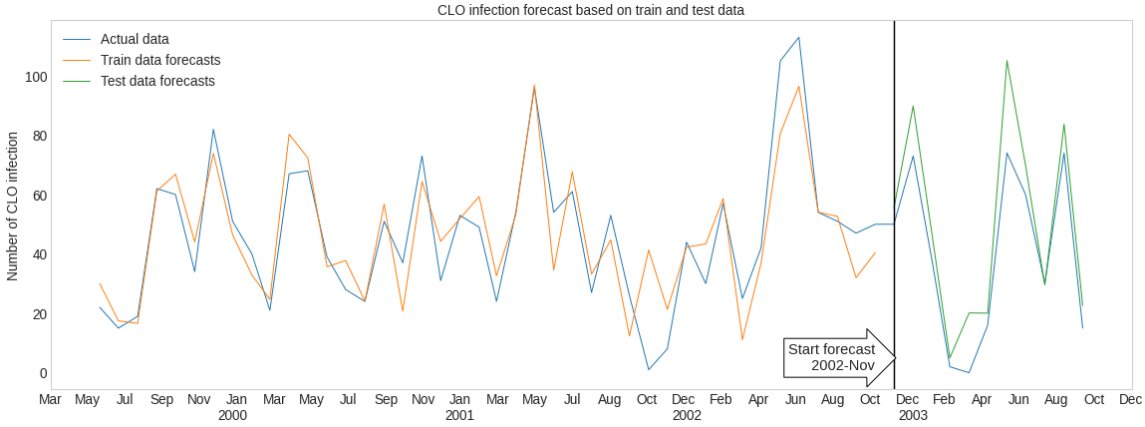


Figure 2. Train and test data prediction with a 90% prediction interval (z=1.645) for MVRM

For predicted data it was calculated lower and upper prediction intervals with 90% probability

$$\text{Upper Prediction Interval } \mathbf{UPI_i} = \hat{y}_i + z\sqrt{\text{MSE}}$$
$$\text{Lower Prediction Interval } \mathbf{LPI_i} = \hat{y}_i - z\sqrt{\text{MSE}} \quad (4)$$

Where:

$$\text{Mean Squared Error (MSE)} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 \quad (5)$$

$$z = 1.645, \text{ the forecasted data with a 90\% prediction interval} \quad (6)$$

Prediction intervals for train data and test data $\mathbf{UPI_i}$ and $\mathbf{LPI_i}$ are represented by green dot line and red dot line, accordingly with 90% probability ($z = 1.645$). (figure 3 and figure 4)

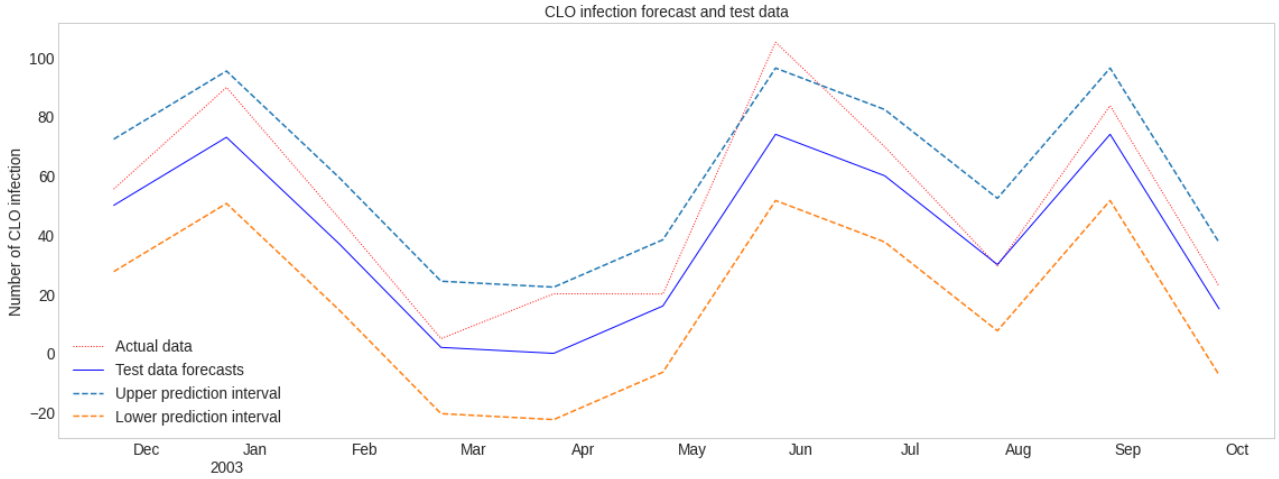


Figure 3. Test data prediction with a 90% prediction interval (z=1.645) for MVRM

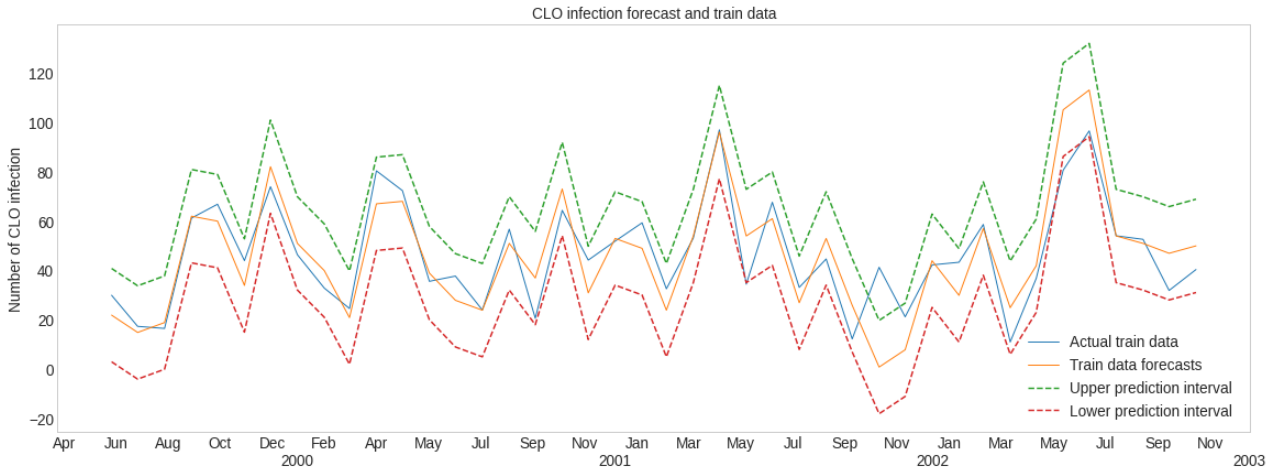


Figure 4. Train data prediction with a 90% prediction interval (z=1.645) for MVRM

4. Conclusions

We suggested new non-linear Multi Varied Regression Model to predict H. pylori infection prevalence based on the data of the Samatya hospital. The new model is constructed by using patterns of H. pylori infection prevalence behaviour in connection with of the mean of humidity, dew point and temperature. Our researched showed that only the forecasting model achieves more accurate results by using the combinations of the given climate variables. The results of this research can help to minimize social cost by predicting the prevalence of the H. pylori infection. The proposed model helps to conduct precise predictive analysis of H. pylori infection prevalence for 1 year based on the dynamics of climate variables. Keeping in mind importance of climate variables in the forecast modelling of H. pylori infection prevalence we found high correlation between the climate factors and the prevalence. This model gives high accurate early forecast results which can be used by hospitals or governments to do early prevention acts against the infection prevalence, since it is critical to save life of people and reduce cost in society.

Acknowledgement.
This research supported by Universiti Malaysia Pahang under projects PGRS200330 and RDU190369

References

[1] Lu, C., Yu, Y., Li, L., Yu, C., & Xu, P. (2018). Systematic review of the relationship of Helicobacter pylori infection with geographical latitude, average annual temperature and average daily sunshine. BMC gastroenterology, 18(1),

[2] Tang, M. Y., Chung, P. H., Chan, H. Y., Tam, P. K., & Wong, K. K. (2019). Recent trends in the prevalence of Helicobacter Pylori in symptomatic children: A 12-year retrospective study in a tertiary centre. Journal of pediatric surgery, 54(2), 255-257.

[3] Peek Jr, R. M., & Blasser, M. J. (1997). Pathophysiology of Helicobacter pylori induced gastritis and peptic ulcer disease. The American journal of medicine, 102(2), 200-207.

[4] Rawla, P., & Barsouk, A. (2019). Epidemiology of gastric cancer: global trends, risk factors and prevention. Przeglad gastroenterologiczny, 14(1), 26.

[5] Chae S, Kwon S, Lee D. Predicting Infectious Disease Using Deep Learning and Big Data. Int J Environ Res Public Health. 2018 Jul 27;15(8):1596. doi: 10.3390/ijerph15081596. PMID: 30060525; PMCID: PMC6121625.

[6] Song, Y., Wang, F., Wang, B., Tao, S., Zhang, H., Liu, S., Ramirez, O., & Zeng, Q. (2015). Time series analyses of hand, foot and mouth disease integrating weather variables. PloS one, 10(3), e0117296.

[7] Forecasting: Methods and Applications, 3rd Edition. Spyros G. Makridakis, Steven C. Wheelwright, Rob J. Hyndman.ISBN: 978-0-471-53233-0 December 1997 656 Pages