IJBPAS, December, 2019, 8(12): 2287-2295

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

FORECASTING POTATO PRODUCTION IN PAKISTAN USING BAYESIAN NON PARAMETRIC MODELLING

KHURRAM H[*1, 2] AND IQBAL MM[3]

1: PhD Scholar, Department of Statistics, Bahauddin Zakariya University, Multan

2: Lecturer, Department of Sciences and Humanities, National University of Computer and Emerging Sciences, Chiniot-Faisalabad

3: Associate professor, Department of Statistics, Bahauddin Zakariya University, Multan

***Corresponding Author: E Mail: haris.khurram@nu.edu.pk**

Received 26th April 2019; Revised 25th May 2019; Accepted 25th June 2019; Available online 1st Dec. 2019

https://doi.org/10.31032/IJBPAS/2019/8.12.4881

ABSTRACT

Forecasting and prediction provide basis for informed decision-making and their importance goes manifold particularly when they have macro impact covering a whole country. Quality of forecast directly affects the governance of a country. Statistics, if applied cautiously, have all the potential to do the job effectively. Forecasting and prediction in relation to the crop production is not new but getting Bayesian approach involved is relatively new and is rare for Pakistani environment let alone Bayesian Non-Parametric, which is the subject matter of the paper, is absent in this part of the world. Life long time-series data of Pakistan as a whole and on provincial basis was used for demonstration. Gaussian Process with an innovative kernel function was devised for the purpose. The performance of the proposed model was compared against existing well-taken models and the supremacy of the proposed model was established.

Keywords: Bayesian Nonparametric Modelling, Gaussian Process, Potato Production, Forecasting

INTRODUCTION

An important aspect of statistical methods is to forecast. There are several procedures and methodologies available in this regard which includes but not limited to Box and Jenkins methodology. The literature on the application of these methodologies to real-life situation is rich. The most-used models are Time Trend, Autoregressive Integrated

Moving Average (ARIMA), and Exponential Smoothing (ES) models. The application of these models is widely available to model and then forecast the agriculture production e.g. [1, 2, 3, 4].

Pakistan is an agriculture country where agriculture is the largest sector of the economy which contributes heavily in country's exports. Forecasting the agriculture yield is an important aspect that helps policymaker and planners. Forecasting of agriculture crop in Pakistan is discussed by [5, 6, 7]. Potato is one of the major agriculture crops which Pakistan is self-sufficient in. The yield has positive growth yearly and is 3975000.73 tons on average for the last three years. As it could earn Forex to Pakistan, we are interest to equip out planners with the crop yield forecast to help good governance.

Bayesian nonparametric models are among leading methods used in various type of data modelling wheremodel has a prior distribution over the regression function in infinite-dimensional space. A famous regression model over infinite-dimensional space with prior knowledge is known to be a Gaussian Process (GP).The application of this model for foreign data modeling is done by [8, 9, 10, 11]. These models are not much popular in Pakistan and the literature is silent on it.

The major objective of this work is to introduce the application of Bayesian Nonparametric domain and to demonstrate its applicability, elasticity and efficiency of such complex models for Pakistani data and to establish its dominance. This approach will open new horizons for new researchers to use, adapt and extend them for similar situations. We used GP approach to forecast the production of potatoes in Pakistan as a whole and to its four provinces separately. Also, a comparison of the Gaussian process with available models is made to highlight its effectiveness and efficiency.

GAUSSIAN PROCESS

A Gaussian Process (GP) is a Stochastic process, indexed by a variable *x*, whose finite distribution is multivariate Gaussian. GP is not a distribution of random variable but a distribution over function with finite domain [13, 10]. Suppose, $x \in X$ is an input space and the mapping $f : X \to \boxed{P}$ is the mapping from input to reals space then $f(x)$ is Gaussian distributed with mean $m(x)$ and covariance function $g(x,x)$, also known as kernel function [12], that is

$$f(x) \boxed{\sim} GP\big(m(x), g(x,x)\big)$$

Now consider a simple model for an output *y* as

$$y = f(x) + e\,,$$

where *e* is a noise term distributed as Gaussian distribution that is $\epsilon \sim N(0, \sigma_n^2)$.

So, the probability distribution of *y* is as

$$p(y/f,x) = GP\left(m(x), g(x,x) + \sigma_n^2 I\right)$$

For a new test data x^* the joint distribution of $f(x)$ and $f(x^*)$ is also Gaussian distribution. The joint distribution of output y for train data and y^* for test data x^* is

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim GP\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} g(x,x)+\sigma_n^2 I & g(x,x^*) \\ g(x^*,x) & g(x^*,x^*) \end{bmatrix}\right).$$

The posterior probability distribution of y^* is also Gaussian distribution

$$p\left(y^*/x^*,x,y\right) = GP\left(m\left(f(x^*)\right), COV\left(f(x^*)\right)\right)$$

Where $m\left(f(x^*)\right) = g(x,x^*)'\left[g(x,x)+\sigma_n^2 I\right]^{-1} y$

and $COV\left(f(x^*)\right) = g(x^*,x^*) - g(x^*,x)\left[g(x,x)+\sigma_n^2 I\right]^{-1} g(x,x^*)$.

The log marginal likelihood is the marginalization of function over function. Thus, the log marginal likelihood is

$$\log p\left(y/x\right) = -\frac{1}{2} y'\left(g(x,x)+\sigma_n^2 I\right)^{-1} y - \frac{1}{2}\log\left|g(x,x)+\sigma_n^2 I\right| - \frac{n}{2}\log(2\pi).$$

The hyper parameters of prior function can be found by optimising the deviance function of the model. In this case the deviance function is the $-2\log p\left(y/x\right)$.

The choice of kernel function for a covariance matrix is one of the important factors need to be considered in GP modelling as it affects the output. The basic properties of a Kernel functions are

a. Kernel function must be symmetric
b. Kernel function must be positive semi definite
c. The sum or product of Kernel functions which satisfy a and b also kernel functions and satisfy a and b

Some general and mostly used kernel functions discussed by [12] are

Squared Exponential Kernel

The squared exponential kernel function is used as by default kernel for GP and for any input x is defined as

$$g_{SE}(x,x) = \sigma_f^2 \exp\left(-\frac{(x-x')^2}{2l^2}\right)$$

Where σ and l are the hyperparameters.

Periodic Kernel

The periodic kernelis used to access the periodic repetition in data. For any input x, the periodic kernel is defined as

$$g_P(x,x) = \exp\left[\frac{-2}{l^2}\mathrm{Sin}\left(\frac{\pi(x-x')}{p}\right)^2\right]$$

Where p and l are hyperparameters.

Linear Kernel

Linear kernel it's self not much important but it provides attractive feature when it is combined with other kernels. So a simple linear kernel is defined as

$$g_L(x,x) = (x-c)(x'-c)$$

METHODOLOGY

The data on potato production in thousand tons for the year from 1948 to 2017 is taken from the official website of http://www.amis.pk. The data is of production of potato separately in all four provinces of Pakistan and finally the total production in Pakistan. Firstly, data is divided into two parts. First part, from 1948 to 2010, was used as training data whereas remaining part, from 2011 to 2017, constituted the second part which was used for testing purpose. Several conventional models are also used to model and forecast the same data under same scheme. The forecast of the conventional model which produced best result was then compared with that of the proposed model by means of certain summary statistics. The one-step-ahead forecasting using GP model for each test data point is also done. This procedure is repeated separately for each province as well as for total production of country.

For the selection of the best ARIMA and ES Model, we calculated the AIC and MAE. Finally, for each test point we measured the Mean Absolute Percentage Error (MAPE) that is

$$\text{MAPE} = \frac{1}{n}\sum \left| \frac{y_{act} - y_{est}}{y_{act}} \right| \times 100$$

Also, Root Mean Square Error (RMSE) was calculated for each test point that is

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum \left(y_{act} - y_{obs}\right)^2}\ .$$

RESULTS AND DISCUSSION

Figure 1 shows the time plot for each province of Pakistan and the total production of Pakistan. Which shows that the production has gradually increased over years for Punjab and KPK. While for Sindh and Baluchistan production has gradually decreased after 2000.

Table 1 shows the model comparison of all provinces of Pakistan and for the total production of Pakistan. From the Linear trend model, ARIMA model of different orders and different ES models the best model with minimum AIC and RMSE is selected in every case.

For Punjab, the best ARIMA model is ARIMA (1,0,1) and best ES model is Browns ES model. For Sindh, the best model among ARIMA is ARIMA (2,1,2) and best ES model is simple ES model. For KPK the ARIMA (0,1,0) and simple ES model appears as best one. For Baluchistan the ARIMA (0,1,1) and simple ES is selected as best models. For Pakistan, ARIMA (1,0,1) and Brown ES pops up as best one. For exponential smoothing the value of smoothing parameters are also mentioned with the model name in the table.

Table 2 shows the results of hyper parameters estimation of GP priors. In the

second column of this table the scheme of local kernel function is given that is used for the estimation of each province. And after mixing the kernel the joint hyperparameters estimate after optimizing the deviance function is shown in table along with deviance function.

Table 3 shows the forecasting performance of each selected ARIMA, ES and GP model. Actual values of 7 years and the forecasted values obtained from each method is also presented here. We calculated RMSE and MAPE for each forecasted observation. Finally, the average value of each measure is calculated which summarises the overall performance of each model.

From Punjab, on average GP is 5% more efficient than ES and 59% more efficient than ARIMA models. For Sindh, GP is 2.1 times more efficient than ES and 95% more efficient than ARIMA. For KPK, GP is 9.8% more efficient than ES and 11% more efficient than ARIMA models. For Baluchistan, GP is approximately 23% more efficient than ES and ARIMA model. For total production of Pakistan, the GP model is 33.8% more efficient than ES and 68% more efficient than ARIMA (**Table 2**).

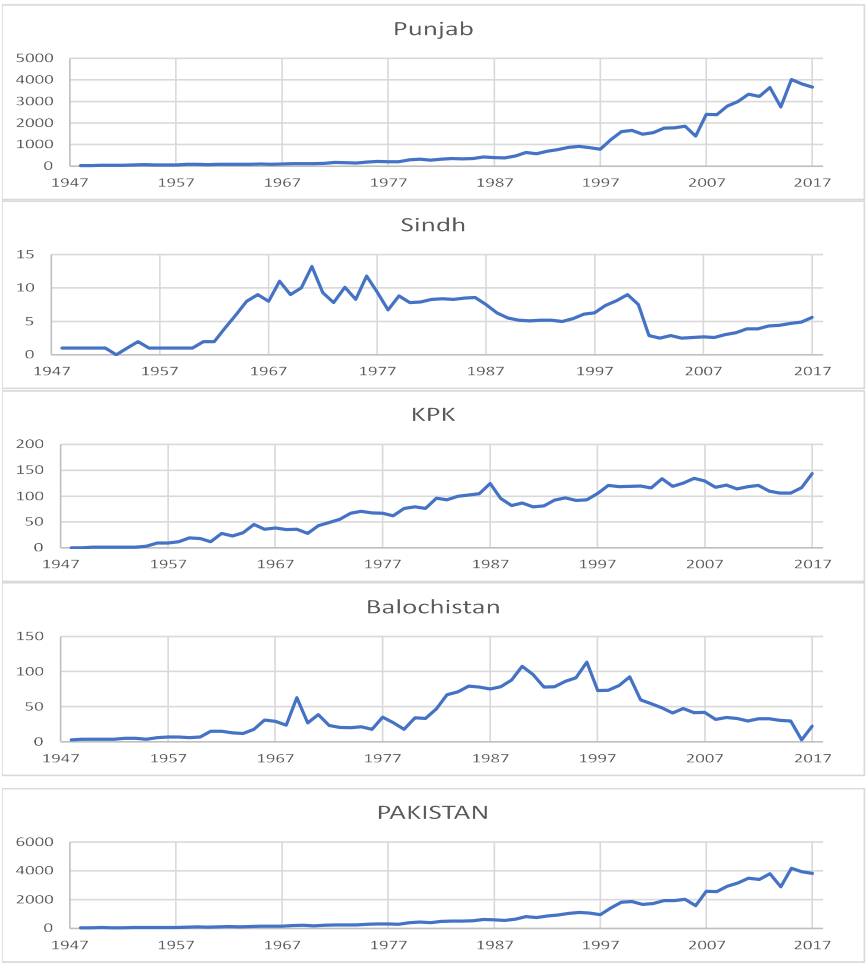

Figure 1: Time plot of the production of potatoes

Table 1: Model Comparisons and Selection for Potato Production

Punjab				Sindh			
Model	**RMSE**	**MAE**	**AIC**	**Model**	**RMSE**	**MAE**	**AIC**
Linear trend $-503.822 + 34.5371\,t$	**406.245**	**326.187**	**12.077**	**Linear trend $4.18126 + 0.04046\,t$**	**3.326**	**2.889**	**2.467**
Simple exponential smoothing $\alpha = 0.8826$	**177.830**	**84.461**	**10.393**	**Simple ES$\alpha = 0.8417$**	**1.437**	**0.962**	**0.756**
Brown's ES$\alpha = 0.3033$	**162.144**	**82.095**	**10.209**	**Brown's ES$\alpha = 0.4251$**	**1.528**	**1.072**	**0.880**
Holt's ES$\alpha = 0.2579$ and $\beta = 0.2373$	**167.037**	**84.771**	**10.300**	**Holt's ES$\alpha = 0.8442$ and $\beta = 0.0094$**	**1.460**	**0.991**	**0.821**
Brown's quadratic ES$\alpha = 0.1487$	**158.004**	**81.914**	**10.157**	**Brown's quadratic ES$\alpha = 0.2989$**	**1.637**	**1.174**	**1.017**
ARIMA(1,0,1)	**139.362**	**68.716**	**9.938**	**ARIMA(2,1,2)**	**1.335**	**0.941**	**0.704**
ARIMA(2,0,1)	**139.270**	**72.652**	**9.968**	**ARIMA(2,1,1)**	**1.374**	**0.930**	**0.731**
ARIMA(1,0,2)	**139.440**	**72.993**	**9.971**	**ARIMA(0,1,0)**	**1.454**	**0.976**	**0.748**
ARIMA(2,0,2)	**140.447**	**72.235**	**10.017**	**ARIMA(1,1,2)**	**1.395**	**0.923**	**0.761**
KPK				**Baluchistan**			
Model	**RMSE**	**MAE**	**AIC**	**Model**	**RMSE**	**MAE**	**AIC**
Linear trend $-7.10584 + 2.30628\,t$	**11.415**	**8.590**	**4.933**	**Linear trend$2.52012 + 1.20577\,t$**	**22.283**	**17.306**	**6.271**
Simple ES$\alpha = 0.8406$	**8.939**	**6.302**	**4.413**	**Simple ES$\alpha = 0.7$**	**11.898**	**7.959**	**4.984**
Brown's linear ES$\alpha = 0.4007$	**9.324**	**6.884**	**4.497**	**Brown's linear ES$\alpha = 0.3288$**	**12.118**	**8.387**	**5.021**
Holt's ES$\alpha = 0.7608$ and $\beta = 0.0329$	**8.852**	**6.297**	**4.425**	**Holt's ES$\alpha = 0.6629$ and $\beta = 0.0687$**	**12.110**	**7.990**	**5.052**
Brown's quadratic ES$\alpha = 0.268$	**9.903**	**7.318**	**4.617**	**Brown's quadratic ES$\alpha = 0.2239$**	**12.626**	**8.904**	**5.103**
ARIMA(0,1,0)	**9.054**	**6.605**	**4.406**	**ARIMA(0,1,1)**	**11.995**	**8.079**	**5.001**
ARIMA(0,1,1)	**9.012**	**6.403**	**4.429**	**ARIMA(1,1,0)**	**12.020**	**8.063**	**5.005**
ARIMA(1,1,0)	**9.018**	**6.383**	**4.430**	**ARIMA(1,0,1)**	**11.973**	**7.942**	**5.029**
ARIMA(1,0,0)	**9.018**	**6.443**	**4.430**	**ARIMA(0,1,0)**	**12.451**	**7.944**	**5.044**
				Pakistan			
Model	**RMSE**	**MAE**	**AIC**	**Model**	**RMSE**	**MAE**	**AIC**
Linear trend$-504.226 + 38.0897\,t$	**385.963**	**310.109**	**11.975**	**ARIMA(1,0,1)**	**140.456**	**68.088**	**9.953**
Simple exponential smoothing $\alpha = 0.867$	**180.090**	**89.389**	**10.419**	**ARIMA(2,0,1)**	**140.628**	**73.496**	**9.987**
Brown's linear ES $\alpha = 0.2896$	**163.826**	**85.846**	**10.229**	**ARIMA(1,0,2)**	**140.895**	**73.128**	**9.991**
Holt'sES $\alpha = 0.2252$ and $\beta = 0.2542$	**167.790**	**84.895**	**10.309**	**ARIMA(2,0,2)**	**143.457**	**72.432**	**10.059**
Brown's quadratic ES $\alpha = 0.1377$	**159.331**	**83.029**	**10.174**	**ARIMA(2,1,2)**	**145.310**	**73.552**	**10.085**

Table 2: Local Kernel functions and estimated Hyperparameters for GP priors

GP	Local Kernel	Deviance	σ_n	σ_f	*l*	*P*	*c*
Punjab	$g_{SE} \times g_P + g_L$	743.4326	17.4509	271.6706	26.3997	5.3945	-56.8611
Sindh	g_{SE}	224.7350	1.3166	6.3225	9.3397	-	-
KPK	$g_{SE} \times g_P \times g_L$	457.8010	8.3398	1.0818	69.6601	-0.3769	-4.0131
Baluchistan	$\sigma_f \times g_P \times g_L$	481.9013	9.3188	1.2605	1.4105	0.9847	-
Pakistan	$g_{SE} + g_P \times g_L$	41496.6900	3.0017	4.0022	2.9997	0.9888	0.0000

Table 3: Forecast performance of the selected model for test observations

	Years	Actual	ARIMA	MAPE	RMSE	ES	MAPE	RMSE	GP	MAPE	RMSE
Punjab	2011	3339.90	3090.31	7.47	24959.00	2987.96	10.54	351.94	3153.81	5.57	186.09
	2012	3235.30	3340.00	3.24	104.70	3173.84	1.90	61.46	3322.38	2.69	87.08
	2013	3639.10	3609.87	0.80	29.23	3365.63	7.52	273.47	3496.83	3.91	142.27
	2014	2743.30	3901.55	42.22	1158.25	3563.33	29.89	820.03	3677.34	34.05	934.04
	2015	4019.90	4216.79	4.90	196.89	3766.94	6.29	252.96	3864.13	3.88	155.77
	2016	3811.10	4557.51	19.59	746.41	3976.46	4.34	165.36	4057.42	6.46	246.32
	2017	3660.30	4925.75	34.57	1265.45	4191.89	14.52	531.59	4257.44	16.31	597.14
	Mean			16.11	535.79		10.71	350.97		10.41	335.53
Sindh	2011	3.90	3.37	13.54	0.53	3.24	16.85	0.66	3.93	0.70	0.03
	2012	3.90	3.31	15.19	0.59	3.24	16.85	0.66	4.29	9.89	0.39
	2013	4.30	3.34	22.43	0.97	3.24	24.59	1.06	4.68	8.74	0.38
	2014	4.40	3.35	23.86	1.05	3.24	26.30	1.16	5.10	15.81	0.70
	2015	4.70	3.31	29.61	1.39	3.24	31.00	1.46	5.54	17.87	0.84
	2016	4.90	3.35	31.59	1.55	3.24	33.82	1.66	6.00	22.39	1.10
	2017	5.60	3.33	40.58	2.27	3.24	42.09	2.36	6.45	15.21	0.85
	Mean			25.26	1.19		27.36	1.29		12.94	0.61
KPK	2011	118.20	113.70	3.81	4.50	114.82	2.86	3.38	116.57	1.38	1.64
	2012	120.60	113.70	5.72	6.90	114.82	4.79	5.78	114.69	4.90	5.91
	2013	109.40	113.70	3.93	4.30	114.82	4.95	5.42	113.50	3.75	4.10
	2014	105.60	113.70	7.67	8.10	114.82	8.73	9.22	114.55	8.47	8.95
	2015	105.60	113.70	7.67	8.10	114.82	8.73	9.22	118.15	11.88	12.55
	2016	116.40	113.70	2.32	2.70	114.82	1.36	1.58	116.02	0.32	0.38
	2017	143.40	113.70	20.71	29.70	114.82	19.93	28.58	119.35	16.78	24.06
	Mean			7.40	9.19		7.34	9.03		6.78	8.22
Baluchistan	2011	29.70	33.85	13.96	4.15	33.85	13.98	4.15	34.35	15.65	4.65
	2012	32.70	33.85	3.50	1.15	33.85	3.52	1.15	31.94	2.34	0.77
	2013	33.10	33.85	2.25	0.75	33.85	2.27	0.75	30.91	6.63	2.19
	2014	30.50	33.85	10.97	3.35	33.85	10.99	3.35	26.93	11.72	3.57
	2015	29.90	33.85	13.19	3.95	33.85	13.22	3.95	29.42	1.61	0.48
	2016	3.00	33.85	1028.17	30.85	33.85	1028.38	30.85	29.51	883.61	26.51
	2017	22.40	33.85	51.09	11.45	33.85	51.12	11.45	29.40	31.23	7.00
	Mean			160.45	7.95		160.50	7.95		136.11	6.45
Pakistan	2011	3491.70	3192.51	8.57	299.19	3106.24	11.04	385.46	3309.62	5.22	182.09
	2012	3392.50	3415.29	0.67	22.79	3278.15	3.37	114.35	3628.49	6.96	235.99
	2013	3785.90	3653.62	3.49	132.28	3455.03	8.74	330.87	3844.63	1.55	58.73
	2014	2883.80	3908.58	35.54	1024.78	3636.89	26.12	753.09	3934.58	36.44	1050.78
	2015	4160.10	4181.33	0.51	21.23	3823.74	8.09	336.36	4167.33	0.17	7.23
	2016	3935.40	4473.12	13.66	537.72	4015.56	2.04	80.16	4065.22	3.30	129.82
	2017	3831.70	4785.26	24.89	953.56	4212.37	9.94	380.67	3717.32	2.99	114.38
	Mean			12.48	427.36		9.90	340.14		8.09	254.15

CONCLUSION

Improvement in Forecasting can help policymakers to makes better decision for future. Forecasting of time series variable has significant importance in literature. In this work is an application of a Bayesian nonparametric model known as GP. There is limited literature available which utilize these models in real life predications and forecasting. Especially in Pakistan where there is a lack of literature related to this field. So, we intended to model the data of potato production of Pakistan and individually its provinces to check the performance of this model. Our results show that the RMSE and MAPE of GP model are found to be lowest as compare to other models. The performance of GP model is more efficient in each case as compared to other methods. This work is a piece of attraction for other reaches to explore the application and versatility of this model in real-life where a simple model is unable to perform well.

REFERENCES

[1] Chandran, K. P., and N. K. Pandey. "Potato price forecasting using seasonal ARIMA approach." Potato Journal 34.1-2 (2007).

[2] Badmus, M. A., and O. S. Ariyo. "Forecasting cultivated areas and production of maize in Nigerian using ARIMA Model." Asian Journal of Agricultural Sciences 3.3 (2011): 171-176.

[3] Taylor, James W. "Short-term electricity demand forecasting using double seasonal exponential smoothing." Journal of the Operational Research Society 54.8 (2003): 799-805.

[4] Jalil, Nur Adilah Abd, Maizah Hura Ahmad, and Norizan Mohamed. "Electricity load demand forecasting using exponential smoothing methods." World Applied Sciences Journal 22.11 (2013): 1540-1543.

[5] Amin, M., M. Amanullah, and A. Akbar. "Time Series Modeling for forecasting wheat production of Pakistan." The Journal of Animal & Plant Sciences 24.5 (2014): 1444-1451.

[6] Iqbal, Najeeb, et al. "Use of the ARIMA model for forecasting wheat area and production in Pakistan." Journal of Agriculture and Social Sciences 1.2 (2005): 120-122.

[7] Masood, M. Asif, Irum Raza, and Saleem Abid. "Forecasting wheat production using time series models in Pakistan." Asian Journal of Agriculture and Rural Development 8.2 (2018): 172-177.

[8] Mojaddady, Mohammad, Moin Nabi, and Shahram Khadivi. "Stock market prediction using twin Gaussian process regression. " International Journal for Advances in Computer Research (JACR) preprint (2011).

[9] Yan, Junchi, *et al*. "Load forecasting using twin Gaussian process model." Proceedings of 2012 IEEE. International Conference on Service Operations and Logistics, and Informatics. IEEE, 2012.

[10] Chen, Zexun. Gaussian process regression methods and extensions for stock market prediction. Diss. Department of Mathematics, 2017.

[11] Chen, Niya, *et al*. "Short-term wind power forecasting using gaussian processes." Twenty-Third Int. Joint Conf. on Artificial Intelligence. 2013.

[12] Williams, Christopher KI, and Carl Edward Rasmussen. Gaussian processes for machine learning. Vol. 2. No. 3. Cambridge, MA: MIT Press, 2006.

[13] Blum, Manuel, and Martin A. Riedmiller. "Optimization of Gaussian process hyperparameters using Rprop." ESANN. 2013.