Cuadernos de Ingeniería. Nueva Serie.
Revista de la Facultad de Ingeniería
de la Universidad Católica de Salta, núm. 10, 2018
ISSN 2545-692X (papel) / 2545-7012 (on line): 29-46

Estimation of the Pearl-Verhulst Logistic function

Estimación de la función logística de Pearl Verhulst

Mika Haarala[1]

Abstract

This paper shows a method to estimate the Pearl-Verhulst function, where we can use a where we can use a linear least squares regression to analyze the set of parameters that is most suitable. We give the mathematical background and two applications: the estimation of the population growth of Salta City and the change of birthrate in Finland.

Keywords: Pearl-Verhulst - logistic function

Resumen

En este artículo se presenta un método para determinar la función de Pearl-Verhulst, haciendo uso de una regresión lineal para analizar el conjunto más adecuado de estimadores más probables. Presentamos sus fundamentos matemáticos y la aplicación a dos casos: la estimación del crecimiento de la población de la Ciudad de Salta (Argentina) y el cambio en la tasa de nacimientos en Finlandia.

Palabras clave: Pearl-Verhulst - función logística

Citar: Haarala, M. (2018). Estimación de la función logística de Pearl Verhulst. *Cuadernos de Ingeniería. Nueva Serie*. [Salta- Argentina], núm.10: 29-46

1 Instituto de Estudios Interdisciplinarios de Ingeniería (IESIING) – Facultad de Ingeniería – UCASAL

Introduction

For some engineering applications, it is necessary to know the population growth of a certain country, city, etc. One quite simple model to do that is the Pearl-Verhulst logistic function (Verhulst, 1838; Pearl and Reed, 1920)

$$N_t = L + \frac{\Delta}{1+e^{at+b}}, \tag{1}$$

where $0 \le L < \min(N_t)$ for all $t \in [-\infty,\infty]$ is a lower limits of the model, $U = L + \Delta > \max(N_t)$ for all $t \in [-\infty,\infty]$ is the upper limit of the model, Δ is the difference between the upper and lower limits, and the parameters $-\infty < a < \infty$, $-\infty < b < \infty$ define the shape of the function. We can also write the function (1) as

$$\log\left(\frac{\Delta}{N_t - L} - 1\right) = at + b.$$

If we could know or estimate the upper and lower limits U and L, then it's possible to estimate the parameters a and b using the least squares estimates as

$$Y = \begin{pmatrix} \log\left(\frac{\Delta}{N_{t_1} - L} - 1\right) \\ \vdots \\ \log\left(\frac{\Delta}{N_{t_n} - L} - 1\right) \end{pmatrix} = \begin{pmatrix} t_1 & 1 \\ \vdots & \vdots \\ t_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = Xu \quad \Leftrightarrow \quad u = \left(X^T X\right)^{-1}\left(X^T Y\right). \tag{2}$$

Definition of the parameters gives

$$\begin{aligned} 0 &< \min(N_t) - L, \\ 0 &< \frac{\Delta}{\max(N_t) - L} - 1 \end{aligned} \tag{3}$$

showing existence of the logarithm, which means that there is a solution (because $t_i \ne t_j$ for all $i \ne j$). Actually, the least square solution (2) means that we are using a regression method under Bayesian conditions for Δ and L. Problem arises when we must find Δ (or U) and L without knowing a and b.

Estimates of the derivative

Differentiation of (1) gives

$$\frac{\partial N_t}{\partial t} = -\frac{a\,\Delta\, e^{at+b}}{\left(1+e^{at+b}\right)^2} = \frac{a\,\Delta}{1+e^{at+b}}\left(\frac{1}{1+e^{at+b}}-1\right).$$

We can see, that the derivative is zero when $at+b=\infty$ (i.e. $a<0, t=-\infty$ or $a>0, t=\infty$) or $at+b=-\infty$ (i.e. $a>0, t=-\infty$ or $a<0, t=\infty$). We can write derivative (3) as

$$\begin{aligned}\frac{a\,\Delta}{1+e^{at+b}}\left(\frac{1}{1+e^{at+b}}-1\right) &= \frac{a}{\Delta}\left(\frac{\Delta}{1+e^{at+b}}+L-L\right)\left(\frac{\Delta}{1+e^{at+b}}+L-L-\Delta\right)\\ &= \frac{a}{\Delta}\left(N_t-L\right)\left(N_t-\left(L+\Delta\right)\right)\\ &= C\left(N_t-L\right)\left(N_t-U\right)\\ &= q_1N_t^2+q_2N_t+q_3.\end{aligned} \tag{4}$$

This shows that the derivative is a parabolic function of N_t with roots L and U ,where q_1, q_2 and q_3 are coefficients to be determined. With these relations deduced from the derivative, we have

$$Y = \begin{pmatrix}\frac{\partial N_1}{\partial t}\\ \vdots \\ \frac{\partial N_n}{\partial t}\end{pmatrix} = \begin{pmatrix} N_1^2 & N_1 & 1\\ \vdots & \vdots & \vdots\\ N_n^2 & N_n & 1\end{pmatrix}\begin{pmatrix}q_1\\ q_2\\ q_3\end{pmatrix} = X\,q \tag{5}$$

whose mean least square solution is

$$q = (X^TX)^{-1}(X^TY). \tag{6}$$

Again, we can use the regression method to estimate the factors of parabolic function.

Since $C=q_1$, we could calculate the parameter a. Because of the estimates of the derivatives are biased, it is better to use the factors q_1, q_2, q_3 only to the calculus of the boundaries and solve the both factors a and b with the new regression as we showed above. This approach comprises of two regressions: firstly we apply the regression for the equation (5) to solve the boundaries and then, these boundaries are used as Bayesian conditions in the regression of the parameters a and b.

Scaling

Normally the matrix $X^T X$ in (6) is close to singular because of large values of N_i. We can define a new variable $u_t = N_t / \lambda$, when the scale of derivatives $\partial u_t/\partial t = \left(\partial N_t/\partial t\right)/\lambda$ is of the orderof $O(1/\lambda)$. Thus, we have a new equation

$$\frac{1}{\lambda}\frac{\partial N_t}{\partial t} = \tilde{q}_1 \frac{N_t^2}{\lambda^2} + \tilde{q}_2 \frac{N_t}{\lambda} + \tilde{q}_3 \tag{7}$$

which yields to

$$\tilde{Y} = \begin{pmatrix} \frac{1}{\lambda}\frac{\partial N_1}{\partial t} \\ \vdots \\ \frac{1}{\lambda}\frac{\partial N_n}{\partial t} \end{pmatrix} = \begin{pmatrix} N_1^2/\lambda^2 & N_1/\lambda & 1 \\ \vdots & \vdots & \vdots \\ N_n^2/\lambda^2 & N_n/\lambda & 1 \end{pmatrix} \tilde{q} = \begin{pmatrix} N_1^2 & N_1 & 1 \\ \vdots & \vdots & \vdots \\ N_n^2 & N_n & 1 \end{pmatrix} \begin{pmatrix} 1/\lambda^2 & & \\ & 1/\lambda & \\ & & 1 \end{pmatrix} \tilde{q} = XD\tilde{q}, \tag{8}$$

where D is a diagonal matrix.

The factor λ can be any large enough number but we are using $\lambda = 10^{\lfloor \log_{10}(\max(N_t)) \rfloor}$, where $\lfloor \cdot \rfloor$ is a floor function which return the maximum integer less or equal than the real value itself. For example, we get $\lfloor 3.6 \rfloor = 3$.

The least square solution of the equation (8) is now in the form

$$\begin{aligned} &(XD)^T XD\tilde{q} = (XD)^T \tilde{Y} \\ \Leftrightarrow \quad &\tilde{q} = \left((XD)^T XD\right)^{-1} \left((XD)^T \tilde{Y}\right) \end{aligned}$$

From this we can see that

$$\begin{aligned} (XD)^T XD\tilde{q} &= (XD)^T \tilde{Y} \\ D^T X^T XD\tilde{q} &= D^T X^T \tilde{Y} \\ DX^T XD\tilde{q} &= DX^T \tilde{Y} \\ D^{-1}DX^T XD\tilde{q} &= D^{-1}DX^T \tilde{Y} \\ X^T XD\tilde{q} &= \lambda^{-1}\ X^T Y \\ \lambda D\tilde{q} &= \left(X^T X\right)^{-1} \left(X^T Y\right) = q\,. \end{aligned} \tag{9}$$

where $D^T = D$ and

$$D^{-1} = \begin{pmatrix} 1/\lambda^2 & & \\ & 1/\lambda & \\ & & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \lambda^2 & & \\ & \lambda & \\ & & 1 \end{pmatrix}.$$

From (9) we see that

$$\begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} = q = \lambda D\tilde{q} = \begin{pmatrix} \tilde{q}_1/\lambda \\ \tilde{q}_2 \\ \lambda\tilde{q}_3 \end{pmatrix}. \quad (10)$$

This is the relation between q and $\tilde{q}$.

Estimation of the parameters

Suppose that $a < 0$. The limits N_∞ and $N_{-\infty}$ can be obtained from (1) as

$$N_{-\infty} = \lim_{t\to-\infty}\left(L + \frac{\Delta}{1+\exp(at+b)}\right) = L$$

$$N_{\infty} = \lim_{t\to\infty}\left(L + \frac{\Delta}{1+\exp(at+b)}\right) = L + \Delta = U$$

Similar way, if $a > 0$, then $N_{-\infty} = U$ and $N_\infty = L$. Thus, the roots of the parabolic equation(4) are the limits as $t \to \pm\infty$. Using weighted variables (10), the parabolic equation (4) is now

$$\frac{\partial N_t}{\partial t} = \frac{\tilde{q}_1}{\lambda} N_t^2 + \tilde{q}_2 N_t + \lambda\tilde{q}_3 = 0,$$

which has roots

$$U = -\frac{\lambda\tilde{q}_2}{2\tilde{q}_1} + \sqrt{\left(\frac{\lambda\tilde{q}_2}{2\tilde{q}_1}\right)^2 - \frac{\lambda^2\tilde{q}_3}{\tilde{q}_1}} = \lambda\left(-\frac{\tilde{q}_2}{2\tilde{q}_1} + \sqrt{\left(\frac{\tilde{q}_2}{2\tilde{q}_1}\right)^2 - \frac{\tilde{q}_3}{\tilde{q}_1}}\right),$$

$$L = -\frac{\lambda\tilde{q}_2}{2\tilde{q}_1} - \sqrt{\left(\frac{\lambda\tilde{q}_2}{2\tilde{q}_1}\right)^2 - \frac{\lambda^2\tilde{q}_3}{\tilde{q}_1}} = \lambda\left(-\frac{\tilde{q}_2}{2\tilde{q}_1} - \sqrt{\left(\frac{\tilde{q}_2}{2\tilde{q}_1}\right)^2 - \frac{\tilde{q}_3}{\tilde{q}_1}}\right).$$

The difference Δ gives

$$\Delta = U - L = \lambda\left(2\sqrt{\left(\frac{\tilde{q}_2}{2\tilde{q}_1}\right)^2 - \frac{\tilde{q}_3}{\tilde{q}_1}}\right).$$

The maximum derivative can be found as

$$N_{t_{\max}} = \frac{U+L}{2} = \lambda\left(-\frac{\tilde{q}_2}{2\tilde{q}_1}\right),$$

when

$$\begin{aligned}\frac{\partial N_{t_{\max}}}{\partial t} &= \lambda\left(\tilde{q}_1\left(-\frac{\tilde{q}_2}{2\tilde{q}_1}\right)^2 + \tilde{q}_2\left(-\frac{\tilde{q}_2}{2\tilde{q}_1}\right) + \tilde{q}_3\right)\\ &= \lambda\left(-\tilde{q}_1\left(\frac{\tilde{q}_2}{2\tilde{q}_1}\right)^2 + \tilde{q}_3\right)\\ &= \lambda\left(-\tilde{q}_1\left[\left(\frac{\tilde{q}_2}{2\tilde{q}_1}\right)^2 - \frac{\tilde{q}_3}{\tilde{q}_1}\right]\right).\end{aligned}$$

The time $t_{\max}$ for the maximum derivative is

$$t_{\max} = \frac{1}{a}\left(\log\left(\frac{\Delta}{\frac{U+L}{2} - L} - 1\right) - b\right) = -\frac{b}{a}.$$

Because of the second derivative is zero at this point, the population rate grows quicker up to this point. So, the rate of population growth increases up until $t_{\max}$ and from this on, it slows down. We call the point as a turning point $t_{\max}$.

We saw above that we can scale the data in a simple way to avoid problems in the numerical calculations. Moreover, we must remember that the time series do not follow the logistic function if

$$\left(\frac{\tilde{q}_2}{2\tilde{q}_1}\right)^2 - \frac{\tilde{q}_3}{\tilde{q}_1} \leq 0.$$

Estimates for derivatives

One key problem to consider is the estimation of derivatives; we can make it in three ways:

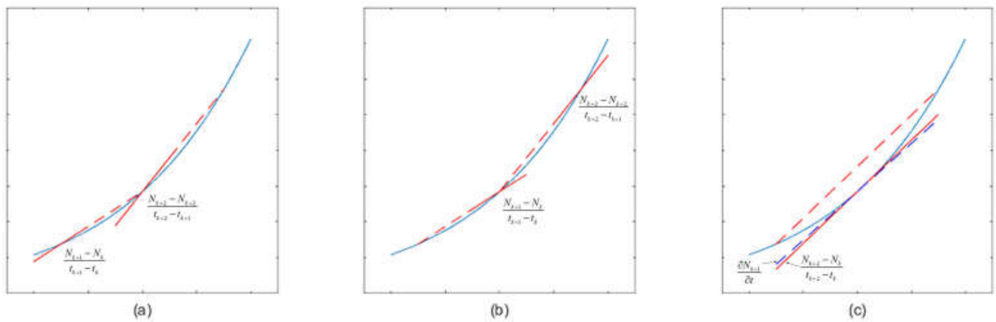

Figure 1: Estimators for the derivative

$$\frac{\partial N_k}{\partial t} = \frac{N_{k+1} - N_k}{t_{k+1} - t_k}, \qquad (a)$$

$$\frac{\partial N_{k+1}}{\partial t} = \frac{N_{k+1} - N_k}{t_{k+1} - t_k}, \qquad (b)$$

$$\frac{\partial N_{k+1}}{\partial t} = \frac{N_{k+2} - N_k}{t_{k+2} - t_k}. \qquad (c)$$

We can see in the Figure 1, that the case (a) overestimates and the case (b) underestimates the real derivative. . The best estimate can be found by using the estimator (c). This is normally recommended in numerical applications (Press et al, 1992). Despite it being the best choice, it still has small bias (systematic error) as we can see in Figure 1 (c); furthermore it cannot use the edge values; this can be a problem when the time series is short. This is the reason why we write the derivatives in (5) as

$$Y = \begin{pmatrix} \dfrac{\partial N_1}{\partial t} = \dfrac{N_2 - N_1}{t_2 - t_1} & \text{at } t_1 \\ \dfrac{\partial N_2}{\partial t} = \dfrac{N_3 - N_1}{t_3 - t_1} & \text{at } t_2 \\ \vdots & \vdots \\ \dfrac{\partial N_{n-1}}{\partial t} = \dfrac{N_n - N_{n-2}}{t_n - t_{n-2}} & \text{at } t_{n-1} \\ \dfrac{\partial N_n}{\partial t} = \dfrac{N_n - N_{n-1}}{t_n - t_{n-1}} & \text{at } t_n \end{pmatrix}.$$

Applications

Population growth in Salta City

The population growth in Salta City, Argentina, starting from the year 1895 is presented in the Table 1 (first row), while the calculated estimates $\tilde{N}_t$ are presented in the second row. We can see in Figure 2 that the curve fit the data quite well.

Table 1: Population growth in Salta City (INDEC)

Year	1895	1914	1947	1960	1970	1980	1991	2001	2010
N_t	20361	33636	76552	123172	182535	265995	373586	472971	536113
$\tilde{N}_t$	21491	27037	71938	123145	185743	269700	377556	473584	546409

To have an idea about the behavior of the model, we will calculate the estimates with different intervals where the starting point is $t_0 = 1895$, 1914, 1947, 1960, 1970 and the end point is $t_1 = 2010$ in all cases. The resulting curves are printed in the Figure 3 and their estimates and statistics can be found from Tables 2 and 3. In these Tables, F is a ratio of the sum of the squares due to regression to the residual mean square (F-test) and R^2 is the coefficient of determination.

Analyzing Table 2, we can see that the estimates of the limits L and U are biased in the cases 1960 and 1970, hence the results. We observe the same phenomenon in the Figure 3, where the last portion of the curve is drawn in detail in the upper left corner.

Tables 2 and 3 show that the best sample to make the estimations is the interval from 1914 until 2010. But the sample 1947-2010 is also good even though its statistics are a little less significant than the sample 1914-2010. The Table 4 shows that the results obtained considering the sample 1947-2010 (third column), match very well with the data of years 1991, 2001, 2010. If we see the Table 3, R^2 gets good values for all sample sizes, except 1895-2010 one.

In order to select one sample to estimate the population growth rate of Salta for the future, we consider that 1947 sample has lower F value than the 1914 and 1960 samples (even though is close to that of 1914) but at the same time, it has the best . Besides, we realize that $L = 15026$ is smaller than the value 20361 reached for the sample 1895 and the standard deviation is 4386 (Table 4). Taking the sample 1947-2010, we estimate the population of Salta City in 2020 as 596048 inhabitants and 636516 inhabitants in 2030. Moreover, the model expects a maximum number of the inhabitants of 700000.

Table 2: Estimates and statistics for the derivative model

Year	L	U	$\tilde{q}_1$	$\tilde{q}_2$	$\tilde{q}_3$	F	R^2
1895	18725	705697	-0.008±0.001 t = -10.87	0.059±0.004 t = 14.14	-0.011±0.004 t = -2.60	170	98.27%
1914	23843	697930	-0.009±0.001 t = -10.25	0.062±0.005 t = 12.69	-0.014±0.005 t = -2.66	125	98.04%
1947	15026	699924	-0.008±0.001 t = -11.84	0.059±0.004 t = 13.58	-0.009±0.005 t = -1.60	122	98.39%
1960	-4362	707532	-0.008±0.001 t = -10.85	0.054±0.005 t = 11.40	0.002±0.007 t = 0.35	68	97.84%
1970	-48841	720040	-0.007±0.001 t = -6.11	0.044±0.008 t = 5.65	0.023±0.013 t = 1.81	23	95.90%

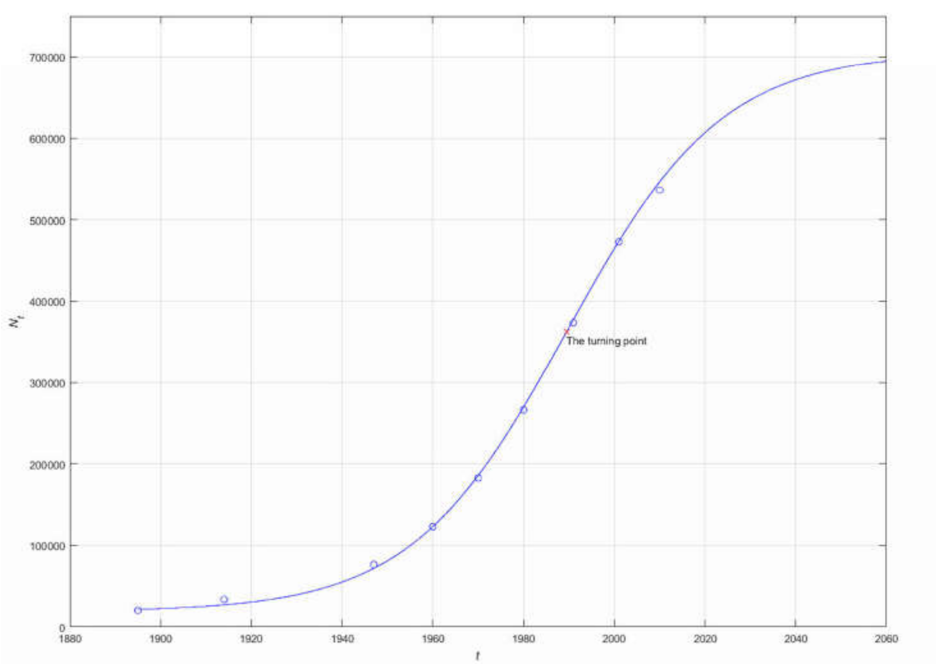

Figure 2: Real and estimated population growth in Salta City

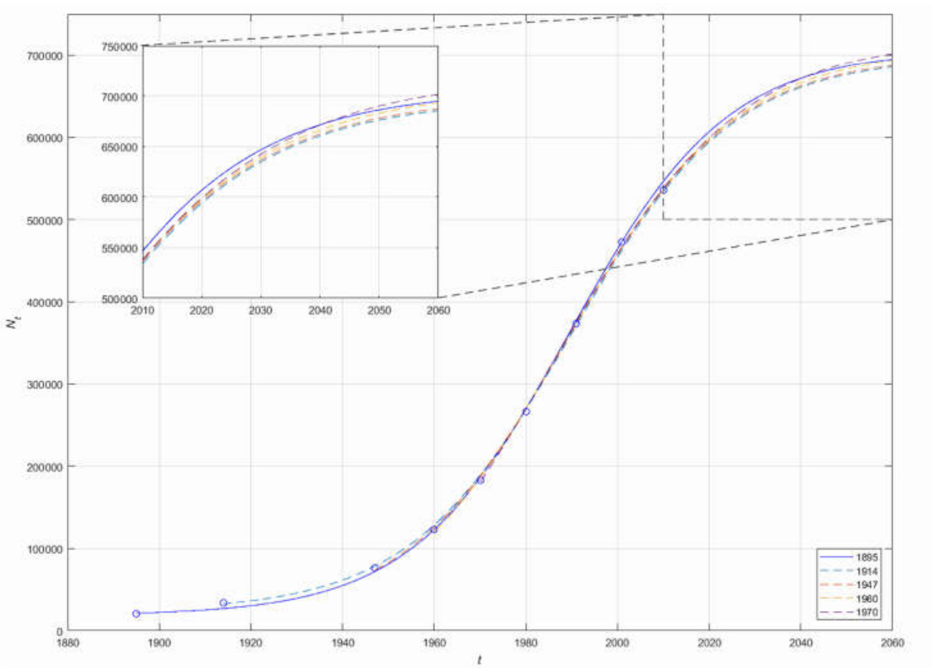

Figure 3: Different estimations for the population growth

Table 3: Bayesian regression results, values of L and U given by Table 2

Year	a	b	F	R^2
1895	-0.058±0.003 t = -21.07	5.51±0.21 t = 25.75	444	98.45%
1914	-0.057±0.001 t = -82.95	4.29±0.04 t = 97.44	6880	99.91%
1947	-0.056±0.008 t = -73.05	2.37±0.03 t = 79.40	5337	99.91%
1960	-0.054±0.001 t = -76.17	1.55±0.02 t = 71.03	5801	99.93%
1970	-0.051±0.001 t = -65.78	0.85±0.02 t = 44.56	4328	99.93%

Table 4: Estimates for the samples 1914-2010 (second row) and 1947-2010 (third row)

Year	1914	1947	1960	1970	1980	1991	2001	2010	2020	2030	*std*
N_t	33636	76552	123172	182535	265995	373586	472971	536113	-	-	-
$\tilde{N}_t$	32919	78509	128653	188856	268760	371209	463206	534088	593951	634654	5165
$\tilde{N}_t$	-	73399	125916	188243	270038	373701	465903	536541	596048	636516	4386

We must remember that these predictions are based on the hypothesis that the population growth rate of last 120 years will remain unchanged. Of course, it is possible that some perturbations of the estimated path occur (for example, lower birthrate, increasing emigration or immigration, etc.).

Birthrate in Finland

In the second application of the model, we will apply the logistic function in the decreasing case. The quarterly data of the birthrate in Finland is tabulated in Table 5. The three first values of 2018 were reported by Tilastokeskus (Statistics Finland).

Table 5: The quarterly birthrate in Finland on 2002-2017 (Tilastokeskus a)
Numbers of the year 2018 are estimated (Tilastokeskus b)

T_i	$Q_1(T_i)$	$Q_2(T_i)$	$Q_3(T_i)$	$Q_4(T_i)$	*Total*
2002	13278	14042	14813	13422	55555
2003	13829	14127	14982	13692	56630
2004	14341	14470	15058	13889	57758
2005	14003	14683	15212	13847	57745
2006	14655	14942	15179	14064	58840
2007	14553	14959	15161	14056	58729

T_i	$Q_1(T_i)$	$Q_2(T_i)$	$Q_3(T_i)$	$Q_4(T_i)$	*Total*
2008	14956	14995	15412	14167	59530
2009	14792	15286	15780	14572	60430
2010	15037	15162	15993	14788	60980
2011	14734	14938	16017	14272	59961
2012	14774	14872	15762	14085	59493
2013	14270	14720	15406	13738	58134
2014	13939	14560	15344	13389	57232
2015	13471	13958	14648	13395	55472
2016	13043	13587	13950	12234	52814
2017	12146	12821	13378	11976	50321
2018*	11644	12094	12438		

We can see from the Figure 4 that each quarterly time series has quite strong variation so they are difficult to model by the Pearl-Verhulst logistic function. . For this reason we give the missing data point as

$$Q_4(2018) = 11976 \times \frac{11644 + 12094 + 12438}{12146 + 12821 + 13378} = 11299.$$

The estimated total sum of the year 2018 (last column in Table 5) is then 47475. The data from the Table 5 including our estimates, are represented in the Figure 4. We can clearly see that all of the curves follow a nonlinear trend. Because of the values of the population size cannot be less than zero, it is quite an attractive idea to apply the Pearl-Verhulst model. Moreover, since we have

$$Q_4(T_i) < Q_1(T_i) < Q_2(T_i) < Q_3(T_i)$$

almost every year except in 2002, the model should be applied to the annual data, to avoid this systematic cycle within the span of a year.

The year when population decrease starts is not so clear, even for 2011 which is the first year to show a decreasing birthrate. We can see from Table 5 that $Q_3(2010) < Q_3(2011)$ and $Q_1(2011) < Q_1(2012)$. If we consider the maxima of each curve, we get the first year for the population decreases over all four quarters. So, we will present our results for the samples starting 2011 and 2012 and the end year of the samples will be 2016, 2017 and 2018. Because we have only estimated numbers in the case of 2018, we will present 3 different estimated results depending upon the variation of the estimates: (a) 47675, (b) 47475 and (c) 47275. The model does not work with sample 2012-2016. The results are shown in Tables 6 and 7 and in Figure 5.

We can see from the Table 6 that the F-value almost double when we add a new event. Also, when the samples are of the same length (for example samples 2011-2016 and 2012-2017 or 2011-2017 and 2012-2018), the F values are similar. The resultant F values are so high that permit us to state that the logistic model fits very well with data.

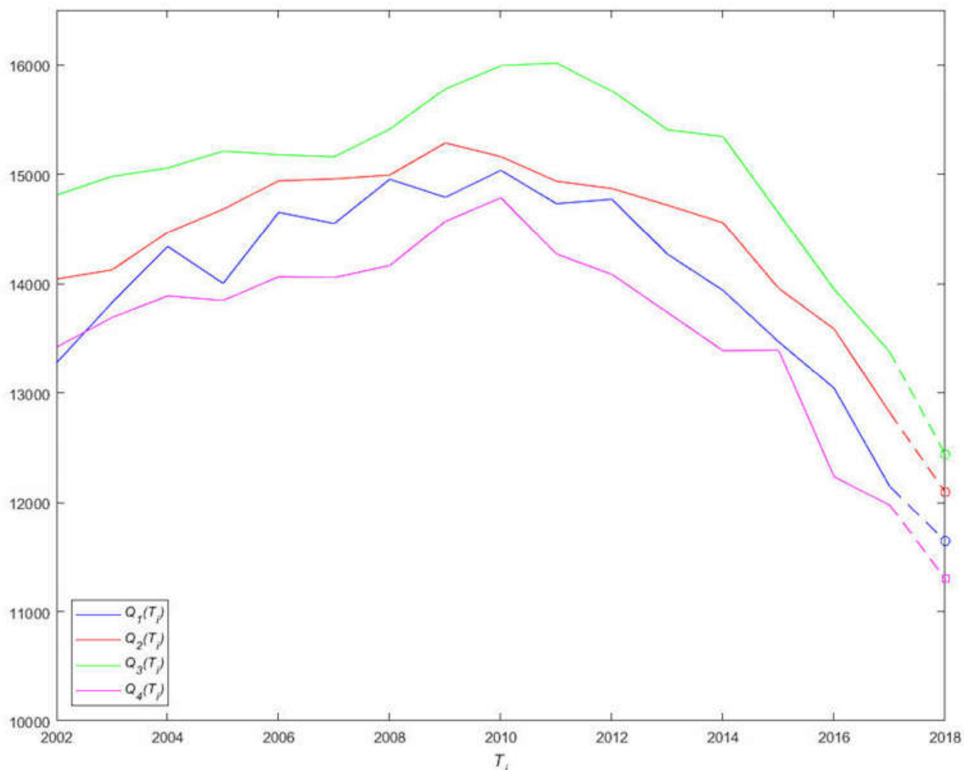

Figure 4: The quarterly birthrate in Finland (2002-2018).

Table 6: Estimates and statistic for the derivative model

Sample	L	U	$\tilde{q}_1$	$\tilde{q}_2$	$\tilde{q}_3$	F	R^2
2011-2016	22946	61372	0.11±0.15 t = 0.71	-0.9±1.7 t = -0.53	1.5±4.8 t = 0.31	45	96.74%
2011-2017	40971	60989	0.26±0.08 t = 3.09	-2.6±0.9 t = -2.85	6.4±2.5 t = 2.54	53	96.37%
2011-2018 (a)	38363	61138	0.21±0.04 t = 4.71	-2.1±0.5 t = -4.33	4.9±1.3 t = 3.80	85	97.14%
2011-2018 (b)	36207	61218	0.18±0.04 t = 4.36	-1.8±0.4 t = -4.36	4.0±1.2 t = 3.33	100	97.55%
2011-2018 (c)	33592	61296	0.16±0.04 t = 4.00	-1.5±0.4 t = -3.52	3.2±1.1 t = 2.85	117	97.90%
2012-2017	32689	62852	0.12±0.18 t = 0.67	-1.1±1.9 t = -0.59	2.4±5.3 t = 0.46	7	82.92%
2012-2018 (a)	33819	62666	0.13±0.09 t = 1.49	-1.2±0.9 t = -1.34	2.7±2.5 t = 1.10	14	87.30%
2012-2018 (b)	29422	62908	0.10±0.08 t = 1.28	-1.0±0.9 t = -1.10	1.9±2.3 t = 0.83	17	89.68%
2012-2018 (c)	23120	63154	0.08±0.08 t = 1.06	-0.7±0.8 t = -0.86	1.2±2.1 t = 0.54	22	91.58%

Table 7: Bayesian regression results for L and U given by Table 6

Sample	a	b	$t_{\max}$	dN_t/dt	F	R^2
2011-2016	0.40±0.02 t = 24.35	-3.29±0.05 t = -65.54	Feb 2020	-3900	593	99.33%
2011-2017	0.51±0.02 t = 33.80	-2.94±0.05 t = -54.14	Oct 2017	-2567	1142	99.56%
2011-2018 (a)	0.47±0.01 t = 45.70	-2.94±0.04 t = -68.04	Mar 2018	-2698	2089	99.71%
2011-2018 (b)	0.45±0.01 t = 45.80	-2.97±0.04 t = -71.97	Jul 2018	-2834	2098	99.71%
2011-2018 (c)	0.43±0.01 t = 45.77	-3.02±0.04 t = -76.16	Dec. 2018	-3007	2095	99.71%
2012-2017	0.34±0.02 t = 23.47	-2.09±0.04 t = -47.29	Feb 2019	-2688	550	99.28%
2012-2018 (a)	0.36±0.01 t = 32.89	-2.11±0.04 t = -53.40	Nov 2018	-2679	1082	99.54%
2012-2018 (b)	0.34±0.01 t = 33.44	-2.19±0.04 t = -60.72	Jul 2019	-2891	1118	99.55%
2012-2018 (c)	0.31±0.01 t = 33.91	-2.31±0.03 t = -69.92	Jun 2020	-3212	1149	99.57%

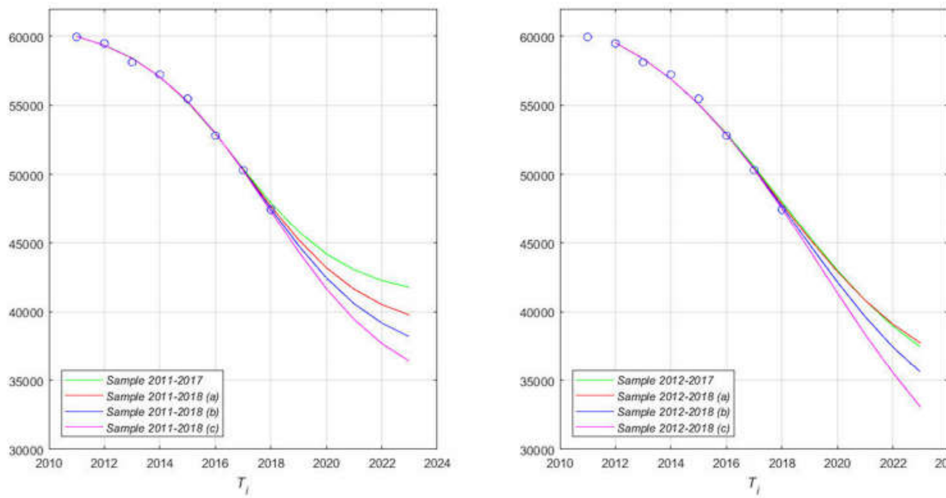

Figure 5: Predictions for the different kind of samples

As Figure 5 and Table 7 show, we get better fit starting the sample from the year 2011. Anyway, the lower limit shows noticeable variations for different samples considered.

The differences between the estimates of years 2019 and 2020 taking the starting points as the years 2011 and 2012, are small, as can be observed in Figure 5. Consequently, it could be difficult to find final estimates before 2021 (official data will be published on 2022). As we mentioned before, the derivative model can have a bias. We will simulate about what would happen if the data follow (A) the path of the sample 2011-2018 (b) or (B) the path of sample 2012-2018 (b). The results are summarized in Table 8.

For the interval 2011-2017 we use the original events (Table 5) and for 2018-2028 the data are estimated following the cases A or B; the columns $N_{t_i}^{(case)}$ present the results; if we consider random perturbations (epsilon) we get the results of columns $N_{t_i}^{(case)} + \varepsilon$, where the *case* is A or B depending on if the lower level (L) is 36207 or 29422, respectively (see Table 7). Using these artificial data, we calculate the lower limits $L_{2011}^{(case)}(T_i)$ and $L_{2012}^{(case)}(T_i)$ using the samples $2011 - T_i$ and $2012 - T_i$, respectively.

We can see that the samples starting from the year 2011 give better estimates than the samples which start from the year 2012. But we also see, that they both must indicate the same lower limit. In both cases, we can have good approximate estimates for the lower limit after 2 years from the turning point. That is to say that we will be able to estimate the lower limit in reliable way in the case A, about 2020 (the turning point is July 2018) and in 2021 in the case B (the turning point is July 2019). It seems that the final lower limit will be between 29400 and 36200.

Our result are trunated at to the estimate of the year 2018. We need to wait some months to get the official numbers because the lower limit especially depends on the event of year 2018. In anyway, we need to wait 2-3 years to get the final reliable estimate, of how low the birthrate will drop in Finland. At the moment, it seems that the birthrate will be 44850 in 2019, 43300 in 2020 and about 40000 in 2021.

Because of the lower limit theoretically takes an infinite amount of time to reach, we can say that the logistic model is invalid when $\left|\tilde{N}_{t_i} - \tilde{N}_{t_i}\right| < std$. That is to say, when the curve is close to the lower limit it follows a linear trend. In anyway, it is quite probable that effects of the current trend in Finland, for which the logistic function is provoking, finish before the curve reaches its minimum. But there is impossible to predict that.

Table 8: The lower limit simulation

T_i	$N_{t_i}^{(A)}$	$N_{t_i}^{(A)}+\varepsilon$	$L_{2011}^{(A)}(T_i)$	$L_{2012}^{(A)}(T_i)$	$N_{t_i}^{(B)}$	$N_{t_i}^{(B)}+\varepsilon$	$L_{2011}^{(B)}(T_i)$	$L_{2012}^{(B)}(T_i)$
2018	47511	47447	35871	28678	47695	47631	37923	32971
2019	44815	44752	35308	32239	44897	44834	34331	30159
2020	42468	42641	36667	35554	42170	42343	33964	31823
2021	40590	40596	35177	34186	39650	39656	29754	27006
2022	39186	39219	35811	35310	37435	37468	29936	28404
2023	38188	38062	35540	35165	35573	35447	28829	27593
2024	37505	37487	36019	35806	34065	34047	29604	28922
2025	37049	37092	36127	35975	32880	32923	29654	29174
2026	36749	36806	36163	36045	31971	32028	29619	29257
2027	36555	36499	36076	35974	31286	31230	29382	29078
2028	36429	36287	36043	35957	30777	30635	29303	29056

Conclusion

We gave a double regression method to analyze the probability of the estimates. The future will show how close our estimates will be to the real values of the problems we considered. As we showed in two examples, it is good to analyze the data, rather than merely taking a sample. In that case, we can find the set of parameters that is most suitable (according to the F-test).

It is very important that the data to be used have few interevent variability («less noise»), because in that way, we would have good estimates for the derivatives. Note that the most difficult step of our technique, is to model the derivatives. If the method passes that step reliably, then we can consider our technique to be applicable for the data.

References

INDEC. Censo Nacional de Población y Vivienda, 1980 y 1991. Censo Nacional de Población, Hogares y Viviendas 2001 y 2010. http://estadisticas.salta.gov.ar/web/archivos/anuarios/anuario2014-2015/A15c020101.xlsx. (Last access: 8.11.2018).

Pearl, R. and L. J. Reed (1920). On the rate of growth of the population of the United States since 1790 and its mathematical representation, *Proc. Nat. Acad. Sci.*, vol. 6, 275-288.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical recipes in FORTRAN: The art of scientific computing*, 2nd ed., Cambridge University Press, Cambridge.

Tilastokeskus (a). Elävänä syntyneet,http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__vrm__synt/statfin_synt_pxt_003.px/table/tableViewLayout2/?rxid=308d0b91-07a7-46c6-807c-b161b034225f,(Last Access 9.11.2018).

Tilastokeskus (b). Väestönmuutokset neljännesvuosittain, http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__vrm__vamuu/statfin_vamuu_pxt_005.px/table/tableViewLayout2/?rxid=6c2b3d86-5c9d-4be3-8fc3-6008576380c4,(Last Access 9.11.2018).

Verhulst, P. F. (1838). Notice sur la loi que la population poursuit dans son accroissement, *Corr. Math. et Phys.* **10**, 113-121.

Recibido: 9/2018
Aceptado: 11/2018