RESEARCH ARTICLE

BON VIEW PUBLISHING

Spatio-Temporal Attributes of Varicella-Zoster Case Number Trends Assist with Optimizing Machine Learning Predictions

David A. Wood[1,*] (iD)

[1]*DWA Energy Limited, UK*

Abstract: The varicella-zoster virus (VZV) (chickenpox) is a problematic infectious disease with regular outbreaks occurring seasonally in most countries. Being able to predict with accuracy the expected number of cases in future weeks based on historical case trend information is an important goal both locally and nationally. Space and time-related attributes extracted from the case number trends for the previous 12 weeks of historical VZV cases recorded in Hungary. These attributes are able to generate reliable predictions for expected VZV cases for multiple weeks ahead. Supervised machine learning (SML) combined with feature selection optimizers can identify combinations of the most effective of 15 local trend time-series attributes supported. These features are complemented with an additional 10 regional trend attributes providing the spatial dimension. The most practical combination of influential trend attributes varies depending on the number of weeks ahead being forecast. SML models are developed using weekly VZV case data (2005–2014) for the regions of Hungary focusing on the region of Komarom-Esztergom (Kom) northwest of Budapest. SML predictions for up to 4 weeks ahead are most strongly influenced by the local time-series attributes including moving averages (MAs) and seasonality components from recently past weeks. However, for predictions further forward (up to 13 weeks) the SML models also exploit regional trend attributes related to recent past rate-of-change in VZV case numbers to provide effective predictions. The proposed trend-attribute method provides more accurate case predictions than the commonly used univariate case-forecasting methods relying on MA and autoregressive integrated moving models. The applied method also provides a means of data mining the most influential trend attributes and the time ranges of their effectiveness. The flexibility and transparency of the technique provide a robust method that could be applied for forecasting short-term epidemiological case numbers associated with other infectious diseases.

Keywords: weeks-ahead VZV case forecasting, spatio-temporal trend attributes, feature importance analysis, optimized machine learning, epidemiological univariate case-trend analysis

1. Introduction

The varicella-zoster virus (VZV), commonly referred to as chickenpox, is a highly infectious disease with global impact. It occurs most prominently in children of primary school age. Its epidemics tend to be seasonal in most countries with peak outbreaks often occurring in winter and spring [1]. VZV cases show declining trends in many countries where vaccination programs have been conducted [2, 3], but the disease has not been eradicated by such efforts. A further complication is that following a childhood VZV infection, the virus remains present and, temporarily, inactive in the nervous system. Unfortunately, this often re-emerges later in life as the herpes zoster virus (shingles) with more serious consequences for older adults [4]. Chickenpox and shingles collectively represent a substantial burden on health authorities' resources in many countries [5]. Hence, methods that provide reliable short-term, VZV case forecasts, locally and nationally [6], from historical case-trend records of past recent weeks are beneficial for resource and response planning.

A range of forecasting techniques are traditionally applied to epidemiological weekly time-series case trends. From the univariate case trend weekly recordings, moving average (MA) and autoregressive integrated moving average (ARIMA) methods are effective for very short-term forecasting (e.g., predicting one week ahead ($t0$)) using t-1 to t-n weekly case records [7]. Supervised machine learning (SML) [8, 9] and deep learning (SDL) models [10, 11] are also proving to be effective for the short-term forecasting of case trends of infectious diseases. An alternative approach to short-term forecasting of time-series data involves decomposing the univariate historical trend to extract trend-specific information. Some studies have achieved this through wavelet decomposition [12, 13]. The recently proposed trend-attribute analysis method applied to energy and environmental time-series datasets [14] is also effective in this regard. It offers a more transparent approach to short-term time-series forecasting with a

*Corresponding author: David A. Wood, DWA Energy Limited, UK. Email: dw@dwasolutions.com

range of data mining capabilities. This study further develops the trend-attribute method applied to provide national weekly case forecasts for Hungary [6] from a published dataset of VZV weekly case trends [15] by considering spatial as well as temporal components of the historical case numbers. This approach provides reliable, more geographically focused, regional weeks-ahead case forecasts incorporating spatially influential case-trend information.

2. Materials and Methods

### 2.1. VZV weekly case trend dataset for Hungarian regions

The University of California, School of Information and Computer Science published dataset [15, 16] compiles weekly VZV case records for Hungary. It provides data compiled at the national level and for the country's regions over a 10-year period. This study evaluates this dataset using trend-attribute analysis. The VZV case data are provided separately for the 19 counties, the capital city, Budapest (Bud; Figure 1), and the nation as a whole. The weekly case records from 2005 to 2014 provide the temporal
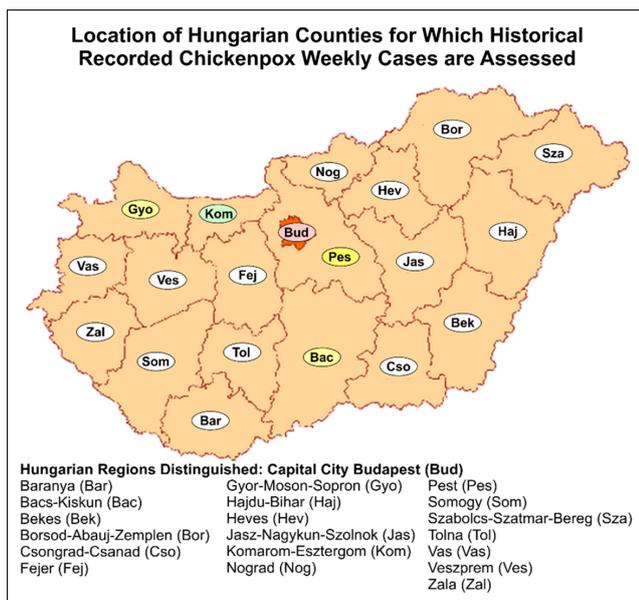
Figure 1. Hungarian regions distinguished are: Capital city Budapest (Bud); Baranya (Bar); Bacs-Kiskun (Bac); Bekes (Bek); Borsod-Abauj-Zemplen (Bor); Csongrad-Csanad (Cso); Fejer (Fej); Gyor-Moson-Sopron (Gyo); Hajdu-Bihar (Haj); Heves (Hev); Jasz-Nagykun-Szolnok (Jas); Komarom-Esztergom (Kom); Nograd (Nog); Pest (Pes); Somogy (Som); Szabolcs-Szatmar-Bereg (Sza); Tolna (Tol); Vas (Vas); Veszprem (Ves); and Zala (Zal). The regions studied for VZV weekly case number data trends to provide short-term forecasts are Kom (highlighted in green) with spatial supporting data from the Bac, Gyo, and Pes regions (highlighted in yellow), together with information from Bud (highlighted in pink) and the case trend for the Hungarian nation as a whole.

dimension. The breakdown of that data for each county and Budapest provides the spatial dimension.

This study focuses specifically on the local case-number trend for the Komarom-Esztergom (Kom; Figure 1) county in northeast Hungary. It is selected for its relatively average population size and VZV recorded case trend compared to other counties as shown in Figure 2 [16]. The Kom time series is challenging from a forecasting perspective in that it displays multi-year periodicity [17] and substantial seasonal swings, with several peak-season highs and lows. Similar to other counties, it displays a clear downward trend in cases recorded, due to an expanding vaccination program from 2005 to 2014. For modeling purposes, the dataset is divided into two subsets: (1) 2005–2011 involving data for 365 weeks used to train and validate the SML models and (2) 2012–2014 involving 157 weeks to independently test the trained and validated models (Figure 2).
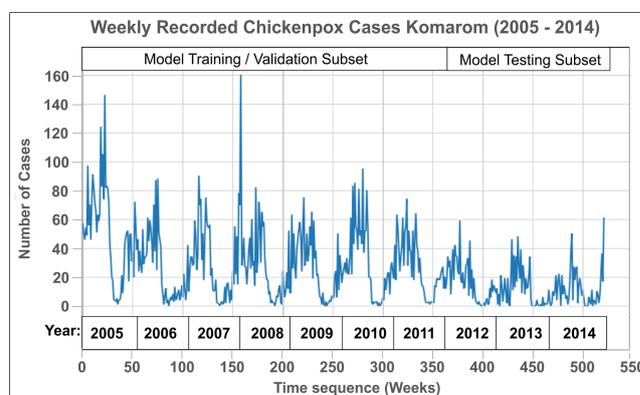
Figure 2. Weekly VZV case trend time series for the Komarom region of Hungary 2005–2014

2.2. Time-series trend attributes calculated

Fifteen trend attributes from the prior 12 weeks ($t-12$ to $t-1$) of recorded data are calculated and appended to each weekly ($t0$) data record for the Kom dataset. These attributes include two seasonal attributes, absolute weekly values, periodic averages, differences, and rates of change. These trend attributes are defined in Table 1. The seasonal attribute (SW) is calculated using the Statsmodels package [18]. The rate of change in that seasonal component between weeks $t-12$ and $t-1$ ($SWdiff$) is then derived from the SW trend. Additionally, two rates of change attributes ($t-6$ to $t-1$ and $t-11$ to $t-1$) are calculated for each of the other Hungarian counties, Budapest, and the Hungarian nation as a whole. These two regional attributes are used to provide spatial input. Based on correlations between these spatially relevant trend attributes defined in Table 1, just 10 additional attributes (Appendix A) are selected as being spatially relevant for the Kom county. These are from counties Bac, Bud, Gyo, Pes, and nationwide, as highlighted in Figure 1. Hence, for the Kom forecasts each compiled data records include 15 local attributes and 10 spatial attributes. These attributes are extracted from the time series, providing the ML models with 25 features to consider.

Table 1. Fifteen trend attributes extracted from the Hungarian VZV case time series

	Weekly case trend attributes for varicella-zoster virus (VZV) cases	Calculation
0	*SW*: Seasonal weekly component	Using Statsmodels [18] algorithm
1	*SWdiff*: Rate of change in *SW* from $t-12$ to $t-1$	$[SW(t-1) - SW(t-12)]/11$
2	*W(t − 1)*: Weekly cases (*W*) for period $(t-1)$	From recorded trend
3	*W(t − 2)*: Weekly cases for period $(t-2)$	From recorded trend
4	*W(t − 3)*: Weekly cases for period $(t-3)$	From recorded trend
5	*AW(t − 1 to t − 3)*: Average weekly cases (AW) from $(t-1)$ to $(t-3)$	$\text{Sum}[W(t-1):W(t-3)]/3$
6	*AW(t − 1 to t − 6)*: Average weekly cases $(t-1)$ to $(t-6)$	$\text{Sum}[W(t-1):W(t-6)]/6$
7	AW(t − 1 to t − 12): Average weekly cases $(t-1)$ to $(t-12)$	$\text{Sum}[W(t-1):W(t-12)]/12$
8	*DW(t − 2 to t − 1)*: Difference in weekly cases (DW) from $(t-2)$ to $(t-1)$	$[W(t-2) - W(t-1)]$
9	*DW(t − 3 to t − 1)*: Difference in weekly cases from $(t-3)$ to $(t-1)$	$[W(t-3) - W(t-1)]$
10	*DW(t − 6 to t − 1)*: Difference in weekly cases from $(t-6)$ to $(t-1)$	$[W(t-6) - W(t-1)]$
11	*DW(t − 12 to t − 1)*: Difference in weekly cases from $(t-12)$ to $(t-1)$	$[W(t-12) - W(t-1)]$
12	*RW(t − 3 to t − 1)*: Rate of change in weekly cases (RW) from $(t-3)$ to $(t-1)$	$[W(t-3) - W(t-1)]/2$
13	*RW(t − 5 to t − 1)*: Rate of change in weekly cases from $(t-5)$ to $(t-1)$	$[W(t-5) - W(t-1)]/4$
14	*RW(t − 8 to t − 1)*: Rate of change in weekly cases from $(t-8)$ to $(t-1)$	$[W(t-8) - W(t-1)]/7$

Note: $t-1$ refers to 1 week before the current week (*t*0); $t-12$ refers to 12 weeks before the current week (*t*0). See also Appendix A for the additional spatially influential trend attributes calculated.

There are four criteria used to select the 15 trend attributes exploited for short-term weeks-ahead forecasting, (1) specific attributes should focus on the recent trends recorded over a previous specified number of weeks (the previous 12 weeks was used as the historical window for this study, which could be extended or shortened based on forecasting results); (2) each attribute should be easy to calculate and visualize. Hence, specific past hourly values, past hourly averages, and rates of change across a range of intervals within the past 12 hours contribute 13 of the 15 attributes selected. (3) The two seasonality-related attributes are slightly more complex to extract but are important because most datasets covering multiple years of historical data display seasonality caused by various influences, which needs to be accounted for. (4) Each selected attribute should exert a distinct influence on potential short-term case-trend forecasts so that collectively multiple distinct influences are brought to bear on specific hours-ahead forecasts. Information about specific input variable influence is typically difficult to extract from more complex attributes such as those derived using wavelet decomposition techniques.

Figure 3 provides a summary workflow diagram describing how the compiled attribute dataset is evaluated with SML and optimizer models. These models conduct feature selection, train, validate, and independently test the SML models. They generate weekly case forecasts for Kom for 5 week-ahead time steps ($t0$, $t+1$, $t+3$, $t+6$, and $t+12$).

2.3. SML methods applied

Four widely used SML models are applied: K-nearest neighbor (KNN) [19], multi-variate Ridge regression [20], support vector regression (SVR) [21], and extreme gradient boosting (XGB) [22]. These SML models applied in Python code using Scikit-learn functions [23]. These models are selected because they apply mathematically distinct methods and can be rapidly executed with relatively few control variables. They have also performed well with the national VZV case trend data from the compiled dataset [6].

The control parameter values, selected by trial and error and grid search [24] and applied to the datasets evaluated, are: (1) KNN – the optimum *K* value varied from 3 to 10 depending on
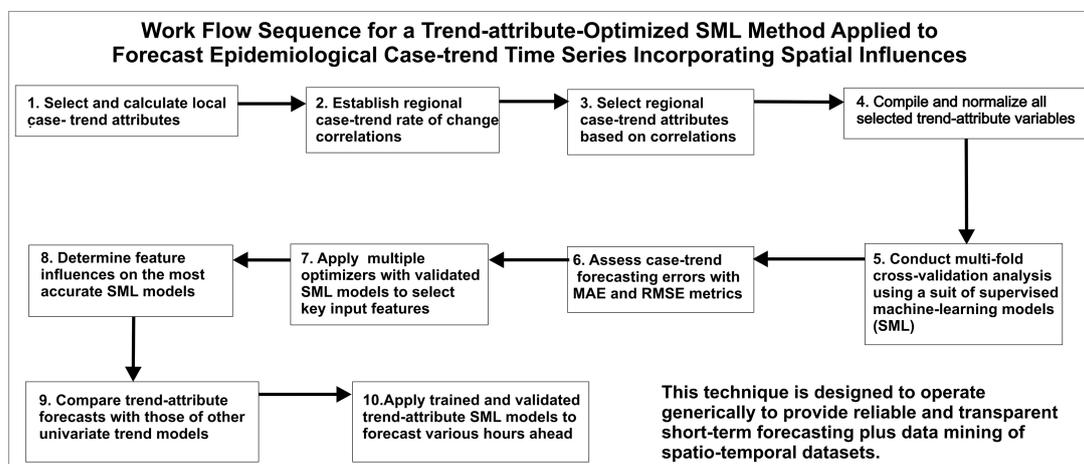
Figure 3. Workflow diagram illustrating how optimized machine learning that feature selects categories of spatio-temporal influence on epidemiological, case-trend time series can be used to provide short-term predictions of contagious diseases in local regions

the dataset and the Manhattan distance measure yielded the best results; (2) Ridge – tolerance factor = 0.001; maximum iterations = 10000; (3) SVR – error-regularization factor (C = 30), an RBF kernel with depth-of-influence factor (γ) set to "scale" so that it is automatically selected based on the variance of the dataset variables, and an error-tolerance limit ($\epsilon = 0.001$); and (4) XGB – number of decision trees (500–750), maximum tree depth (10–15), learning rate η (0.01), subsample limit of data records sampling for each decision tree (0.4–0.6), and the fraction of features (columns) to be considered by each tree (0.8–0.9). Prior to executing these four SML models, the data variables are all normalized to value ranges between −1 and +1 to avoid scaling biases. The forecasting performance of each model solution is then assessed using statistical error metrics mean absolute error (MAE) and root mean squared error (RMSE).

2.4. Multi-K-fold cross-validation

Cross-validation analysis [25] rotates through a random selection of training and validation data subsets. It does this by applying multiple repeat runs with distinct data record combinations to provide statistical confidence in terms of mean and standard deviations of error metrics. The recently developed multi-K-fold cross-validation technique [26] extends basic cross-validation to repeatedly perform such "leave-one-set-out" analysis with a sequence of different percentage splits (e.g., 3-fold in each run rotates three times through 1/3 of the data records randomly assigned to validation and 2/3 assigned to training; 15-fold rotates 15 times through 1/15 of the data records assigned to validation and 14/15 assigned to training). In this study, 3-fold, 4-fold, 5-fold, 10-fold, and 15-fold analysis was conducted and repeated in multiple leave-one-set-out trials. These trials identify the most effective training: validation subset splits and the forecasting uncertainty to be expected from each model. The 4-fold cross-validation method provided the most statistically reliable results for the datasets evaluated in this study. The multi-K-fold analysis is conducted using customized Python-coded Scikit-learn functions "RepeatedKFold" and "cross_val_score" [27].

2.5. Optimized feature (trend-attribute) selection

The KNN algorithm combined separately with five optimizer algorithms applied to the training/validation subset provides an effective way of conducting feature selection. It rapidly identifies the most influential of the 25 compiled trend-attribute combinations for the Kom region. The selected feature combinations are then used with the SML models to provide short-term VZV case number predictions. The five optimizers employed are: differential evolution (DE) [28], cuckoo-search (CSO) [29]; Jaya [30], particle swarm optimizer (PSO) [31], and sin-cosine algorithm (SCA) [32]. As these optimizers all apply different mathematical algorithms to search the feasible solution space, using them collectively tends to identify a wider range of potential feature combinations. The KNN-optimizer models were executed in this study with two sets of attributes available: (1) using only the 15 attributes calculated for the local Kom time series (i.e., no spatial input) and (2) using the full set of 25 compiled trend attributes including both those local to Kom and those extracted from spatially relevant time series (from the Bac, Gyo, and Pes counties and the Hungarian national data).

2.6. Relative feature influence on the SML predictions

Having identified the most effective feature selections and trained/validated the SML models with the high-performing feature combinations, it is informative to consider the relative contributions of the attributes to the specific solutions generated by the SML models. Two of the SML models evaluated readily reveal feature importance information. The Ridge model yields this information from the relative absolute magnitudes it assigns to its regression coefficients. The XGB model yields this by comparing Gini coefficients [33], or its other node selection discriminators, associated with each decision-tree branch in its optimum solutions. In this study, feature influences derived from the XGB model are displayed in bar charts. This information provides analysis and insight that reveals the most influential attributes in the forecasts derived for each time step. This information is also useful for assessing the likelihood that attributes calculated in various ways will make meaningful contributions to specific timestep forecasts.

3. Statistical Relationships Between the Kom Extracted Trend Attributes

3.1. Correlations between trend attributes and time-series case sequences

Pearson (R) and Spearman (p) correlation coefficients reveal the strength of the relationships between each of the 25 attributes calculated for assessing the Kom VZV case time series, and each of the hours-ahead time steps considered. Comparing R and p values also identifies whether the relationships are parametric (an assumption made by the R calculation but not the p calculation) or non-parametric. If R and p values are similar, the variable distributions are approximately parametric. Figure 4 displays the R and p data in the form of heat maps for the 25 attributes distinguished as 15 local attributes (#0 to #14) and 10 spatial attributes (#15 to #24) (defined in Appendix A).

Although the overall R and p relative relationships are similar, the actual values are typically distinct. This suggests that most of the attributes exhibit non-parametric relationships with the time series for each time step considered. Overall, attributes 0–7 (Table 1) show strong positive correlations for time steps $t0$, $t+1$, and $t+3$. However, those correlations weaken progressively as the time step moves forward from $t0$ to $t+12$. For local attributes #8 to #14 and regional attributes #15 to #24, p values tend to be substantially higher than R values. The strongest correlations are displayed for time steps $t+1$, $t+3$, and $t+6$, with weaker correlations for $t0$ and $t+12$.

To be clear, neither R nor p values are used by the SML models when making their predictions. These correlation statistics therefore have no influence on the models' prediction results. The correlations are calculated purely to characterize the variable distribution relations and show the manner in which the trend-attribute variables are correlated with the VZV case number trends.

3.2. Correlation relationships case trends for Hungarian regions

R and p values are also used to assess the Hungarian regional and national VZV case-trend data for potential spatial influences. This is conducted in terms of the two rate-of-change attributes

Attribute Code		Pearson's Correlation Coefficients (R)					Spearman's Correlation Coefficients (R)				
		Case-number time series for Komarom (2005-2014)					Case-number time series for Komarom (2005-2014)				
		t0	t+1	t+3	t+6	t+12	t0	t+1	t+3	t+6	t+12
Komarom Time-series Attributes	0	0.6651	0.5913	0.5052	0.3775	0.0360	0.7361	0.6857	0.6082	0.4387	-0.0272
	1	0.3958	0.4295	0.4565	0.4205	0.2711	0.4166	0.4560	0.5278	0.5182	0.3350
	2	0.6447	0.6893	0.5980	0.3748	0.0309	0.7474	0.7500	0.6476	0.4421	0.0351
	3	0.6922	0.5877	0.4784	0.3536	-0.0343	0.7513	0.6944	0.5695	0.3617	-0.0324
	4	0.5909	0.6016	0.4593	0.2641	0.0043	0.6959	0.6506	0.5243	0.2974	-0.0314
	5	0.7283	0.7095	0.5798	0.3749	0.0003	0.8005	0.7646	0.6287	0.3954	-0.0216
	6	0.6940	0.6404	0.5094	0.3151	-0.0556	0.7402	0.6791	0.5313	0.3011	-0.0894
	7	0.5629	0.5022	0.3701	0.1832	-0.1205	0.5529	0.4817	0.3314	0.1242	-0.1799
	8	-0.0628	0.1163	0.1391	0.0224	0.0778	-0.0409	0.0623	0.1015	0.0897	0.0739
	9	0.0602	0.1032	0.1714	0.1379	0.0338	0.1795	0.0901	0.1533	0.1728	0.0767
	10	0.1688	0.2890	0.3170	0.2148	0.1259	0.2217	0.2994	0.3551	0.3107	0.1592
	11	0.3662	0.4456	0.4119	0.3262	0.1513	0.4604	0.5322	0.5226	0.4431	0.1861
	12	0.0602	0.1032	0.1714	0.1379	0.0338	0.0327	0.0901	0.1533	0.1728	0.0767
	13	0.1712	0.2462	0.2591	0.2066	0.1385	0.1881	0.2610	0.3167	0.3055	0.1690
	14	0.2376	0.3650	0.3556	0.3150	0.1519	0.3163	0.3949	0.4188	0.3964	0.1779
Regional Time-series Attributes	15	0.1246	0.1508	0.1589	0.1860	0.2111	0.1648	0.1968	0.2694	0.3176	0.2727
	16	0.2711	0.2825	0.3521	0.3534	0.2754	0.3541	0.3798	0.4520	0.4534	0.3367
	17	0.1052	0.1481	0.2074	0.2293	0.2051	0.1854	0.1949	0.2921	0.3285	0.2678
	18	0.2788	0.3046	0.3649	0.4007	0.2609	0.3598	0.3871	0.4685	0.4839	0.3291
	19	0.1381	0.1404	0.2262	0.2146	0.1236	0.1810	0.2184	0.3269	0.3252	0.1803
	20	0.3093	0.3582	0.4103	0.3348	0.1373	0.4050	0.4509	0.5407	0.4509	0.1814
	21	0.1321	0.1382	0.2106	0.2198	0.1306	0.1971	0.2244	0.3236	0.3482	0.1773
	22	0.3004	0.2942	0.3792	0.3432	0.1877	0.3890	0.4028	0.5024	0.4514	0.2323
	23	0.1399	0.1828	0.2520	0.2867	0.2689	0.2146	0.2362	0.3692	0.4144	0.3264
	24	0.3205	0.3320	0.4184	0.4388	0.3198	0.3878	0.4065	0.5287	0.5450	0.3939

Figure 4. Heat map of Pearson correlation coefficients (*R*) between trend attributes and Komarom weekly VZV time series (2005–2014)

Regions/ Counties	Correlation Coefficients for Rates of Change in Case Numbers Over the Prior 5-week Interval				Regions/ Counties	Correlation Coefficients for Rates of Change in Case Numbers Over the Prior 10-week Interval			
	Pearson's (R)			Spearman's (p)		Pearson's (R)			Spearman's (p)
	Budapest	Komarom	Nationwide	Nationwide		Budapest	Komarom	Nationwide	Nationwide
Budapest	1.0000	0.4587	0.7741	0.7608	Budapest	1.0000	0.6119	0.8503	0.8473
Baranya	0.3159	0.3611	0.5374	0.6130	Baranya	0.5501	0.5038	0.7037	0.7408
Bacs	0.4297	0.3380	0.5714	0.5897	Bacs	0.5749	0.5551	0.6956	0.7014
Bekes	0.3499	0.3367	0.4795	0.5345	Bekes	0.4640	0.4032	0.5717	0.6535
Borsod	0.4855	0.2840	0.5920	0.6130	Borsod	0.6225	0.4450	0.7122	0.7509
Csongrad	0.3461	0.2025	0.4357	0.5006	Csongrad	0.4756	0.3437	0.5933	0.6465
Fejer	0.3711	0.2776	0.6070	0.6239	Fejer	0.5445	0.4612	0.7283	0.7298
Gyor	0.4341	0.4553	0.6271	0.6107	Gyor	0.5966	0.5917	0.7564	0.7469
Hajdu	0.4478	0.4214	0.6033	0.5789	Hajdu	0.6038	0.5202	0.7196	0.6997
Heves	0.4186	0.3491	0.6163	0.6063	Heves	0.5104	0.4878	0.6796	0.6693
Jasz	0.3797	0.3745	0.5993	0.6140	Jasz	0.5217	0.5310	0.6984	0.6885
Komarom	0.4587	1.0000	0.6298	0.5612	Komarom	0.6119	1.0000	0.7185	0.6855
Nograd	0.3736	0.2845	0.5343	0.4932	Nograd	0.5457	0.4054	0.6676	0.6692
Pest	0.5123	0.5020	0.7469	0.7278	Pest	0.7017	0.6165	0.8501	0.8466
Somogy	0.3635	0.2531	0.5243	0.6112	Somogy	0.4987	0.3454	0.6283	0.7027
Szabolcs	0.2576	0.3129	0.4523	0.4403	Szabolcs	0.4220	0.4354	0.5765	0.6301
Tolna	0.3458	0.2031	0.4588	0.4700	Tolna	0.4867	0.3737	0.5792	0.5943
Vas	0.2729	0.2332	0.4119	0.4855	Vas	0.4382	0.4255	0.5774	0.6562
Veszprem	0.3003	0.3163	0.5683	0.5747	Veszprem	0.5099	0.4037	0.6963	0.7583
Zala	0.1450	0.2031	0.3277	0.4336	Zala	0.4257	0.3519	0.5316	0.6172
Nationwide	0.7741	0.6298	1.0000	1.0000	**Nationwide**	0.8503	0.7185	1.0000	1.0000

Figure 5. Heat map of correlation coefficients between rates of change in VZV cases recorded at regional and national level in Hungary (2005–2014)

($t-6$ to $t-1$ and $t-11$ to $t-1$) calculated for each region. These are expressed as heat maps in Figure 5.

The Komarom (Kom) region (Figure 5) has strong *R* and *p* values for both rate of change attributes with the nationwide data trends, and trends for Budapest (Bud), Pest (Pes), Gyor (Gyo), and Bacs (Bac) regions. It is because of these high correlation coefficient values that trend attributes from those four regions, and the nationwide trends have been included in the datasets (as trend attributes #15 to #24) compiled to forecast the Kom time series. This correlation analysis is therefore a crucial step in establishing which regional data to include as spatial-dimension features for each specific region.

4. Results

A provisional multi-fold cross-validation analysis was conducted with two *t*0 datasets (15-variable dataset with local attributes and 25-variable dataset with local plus spatially relevant

regional attributes) applying multiple SML models. This analysis considered means and standard deviations of the MAE and RMSE error metrics. The four-fold cross-validation, repeated eight times to provide 32 random selections, provided the most statistically reliable results. The four-fold analysis outperformed the 3-fold, 5-fold, 10-fold, and 15-fold cross-validation trials. It established that the KNN, Ridge, SVR, and XGB models all generated reliable predictions with comparable error levels. For the 15-variable dataset, MAE mean varied between 10.563 and 11.170 cases with standard deviations varying from 0.998 to 1.064 cases. For the 25-variable dataset, MAE mean varied between 10.743 and 11.373 cases with standard deviations varying from 1.028 to 1.180 cases. The strong performance of the four-fold cross-validation justified the use of 0.75 training: 0.25 validation splits of the dataset for more detailed SML modeling.

4.1. Optimized feature selection for detailed SML predictions

The KNN model was applied, combined with the five optimizers, to train and validate the ML models (using 2005–2011 data records). This was performed separately for each of the $t0$, $t + 1$, $t + 3$, $t + 6$, and $t + 12$ time series for the 15-variable and 25-variable datasets. Considering the best solutions from multiple KNN-optimizer runs for each time-series dataset, the two best feature combinations were selected for detailed SML analysis for the 15-attribute selection (Table 2) and the 25-attribute selection (Table 3).

Table 2. Feature selections established by KNN-optimizer models from Kom trend attributes only

Features selected from 15-variable data selection		
Case identifier	**Optimizers**	**Features selected**
$t0$ **dataset**		
KOMt0FS3_4Var	Jaya	[0 7 11 14]
KOMt0FS4_10Var	Jaya	[0 2 3 4 6 8 10 11 13 14]
$t + 1$ **dataset**		
KOMt1FS3_10Var	Jaya/PSO	[0 1 4 5 9 10 11 12 13 14]
KOMt1FS4_6Var	CSO	[1 4 5 11 13 14]
$t + 3$ **dataset**		
KOMt3FS3_4Var	Jaya	[1 5 11 13]
KOMt3FS4_5Var	DE	[0 1 5 6 11]
$t + 6$ **dataset**		
KOMt6FS3_5Var	PSO	[0 1 6 7 9]
KOMt6FS4_4Var	SCA	[0 1 6 12]
$t + 12$ **dataset**		
KOMt12FS3_6Var	Jaya/DE	[1 2 7 8 9 12]
KOMt12FS4_8Var	CSO	[1 2 6 7 9 10 12 13]

Note: For feature numbers 0–14 see (Table 1), and for feature numbers 14–24 see Appendix A.

The optimized feature selections reveal the importance of the seasonality attributes (0 and/or 1) to all the solutions. All of the optimized solutions from the 25-variable selection include at least three attributes from the other spatially relevant areas (i.e., attributes 15–24) with attributes 20, 22, and/or 24 appearing in many of them. The optimum solutions derived from the 15-variable selections involve from 4 to 10 features (Table 2). This indicates that for many of those selections more than half of the available features are disregarded. The same is true for the optimum solutions derived from the 25-variable selections, which involve from 7 to 14 features (Table 3).

Table 3. Feature selections established by KNN-optimizer models from Kom plus spatial trend attributes

Features selected from 25-variable data selection		
Case identifier	**Optimizers**	**Features selected**
$t0$ **dataset**		
KOMt0FS1_13Var	Jaya	[0 1 2 4 6 7 8 14 16 17 18 19 24]
KOMt0FS2_10Var	CSO	[0 4 7 8 10 16 17 19 22 24]
$t + 1$ **dataset**		
KOMt1FS1_7Var	CSO	[1 5 11 14 17 18 19]
KOMt1FS2_14Var	Jaya	[0 1 3 5 6 9 10 11 14 16 17 18 20 21]
$t + 3$ **dataset**		
KOMt3FS1_9Var	PSO	[0 1 5 7 11 15 19 20 21]
KOMt3FS2_8Var	DE	[0 1 5 11 15 20 22 23]
$t + 6$ **dataset**		
KOMt6FS1_8Var	CSO	[0 1 3 6 13 20 22 24]
KOMt6FS2_10Var	DE	[0 1 3 6 11 13 15 20 22 24]
$t + 12$ **dataset**		
KOMt12FS1_9Var	PSO	[1 7 9 13 15 20 22 23 24]
KOMt12FS2_12Var	Jaya	[1 4 5 6 7 9 10 13 15 19 20 24]

Note: For feature numbers 0–14 see (Table 1), and for feature numbers 14–24 see Appendix A.

4.2. Four-fold cross-validation of trained SML models for $t0$ to $t + 12$ forecasts (2005–2011)

Four-fold cross-validation analysis was performed with each of the optimized feature selections described in Tables 2 and 3. This was performed applying each of the KNN, Ridge, SVR, and XGB models separately (Figure 6). Hence, for each time-series dataset, six distinct feature combinations were evaluated: 15 variables, FS3 and FS4 selections considering local Komarom information only, and 25 variables, FS1 and FS2 selections considering local Komarom plus spatially relevant information from other areas. The feature configurations generating the lowest mean errors for each timestep sequence are presented in Table 4. The SVR models provide the best forecasts for the validation subsets relating to the $t0$ and $t + 1$
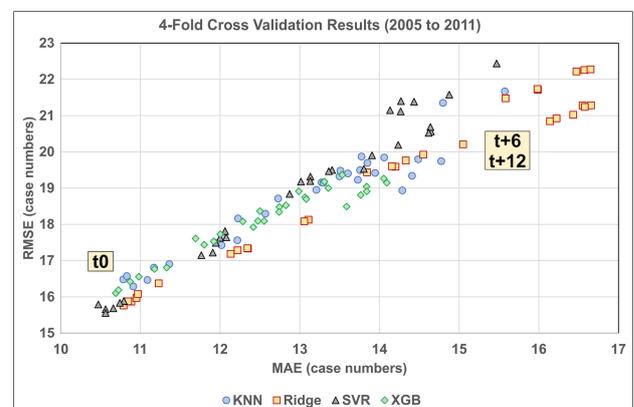
Figure 6. Cross-validation prediction errors RMSE versus MAE for SML models applied to the training and validation subsets (2005–2011) for time series $t0$, $t + 1$, $t + 3$, $t + 6$, and $t + 12$

Table 4. Four-fold cross-validation prediction errors for evaluated ML models

30 trials run for each model		MAE (4-fold CV)		RMSE (4-fold CV)	
Model	Time series	Mean	Standard deviation	Mean	Standard deviation
SVR Opt FS2 (10 variables)	$t0$	10.563	1.098	15.554	2.048
SVR Opt FS2 (14 variables)	$t + 1$	11.765	1.180	17.152	2.500
XGB Opt FS4 (5 Variables)	$t + 3$	12.417	1.092	17.923	2.237
XGB Opt FS4 (4 Variables)	$t + 6$	12.741	1.187	18.342	2.458
XGB Opt FS1 (9 Variables)	$t + 12$	13.589	1.225	18.492	2.571

Cross-validation (CV) results for models achieving the lowest prediction errors for each time series trained and validated (2005–2011)

Note: Value units are VZV case numbers.

Table 5. Independent testing subset prediction errors for evaluated SML models

Testing subset results for models with lowest prediction errors for each time series forecasted (2012–2014)

Model	Time series	MAE	RMSE
KNN – 25 variables	$t0$	9.043	12.031
SVR Opt FS4 (6 variables)	$t + 1$	8.348	10.867
SVR Opt FS4 (5 Variables)	$t + 3$	10.316	13.571
SVR Opt FS2 (10 Variables)	$t + 6$	14.685	17.603
SVR Opt FS2 (12 Variables)	$t + 12$	19.007	21.975

Note: Value units are chickenpox case numbers.

time steps, whereas the XGB models provide the best forecasts for the validation subsets relating to the $t + 3$, $t + 6$, and $t + 12$ time steps. The mean MAE and RMSE values increase progressively from $t0$ to $t + 12$. However, in general, those error values and the associated standard deviations are relatively low.

The cross-validation results of all 120 evaluations (30 for each SML model) are displayed in Figure 6. It is apparent that all SML models provide good forecasts for the validation subsets for timesteps $t0$, $t + 1$, and $t + 3$. However, for time steps $t + 6$ and $t + 12$ there is more dispersion. The SVR and Ridge models are generating more errors for all feature combinations with the validation subsets.

4.3. Application of trained models to independent testing time sequence (2012–2014)

The trained and validated models based on 2005–2011 data were all applied to forecast the testing dataset (2012–2014). The feature configurations and SML models that generated the lowest errors for each time step with the testing dataset are presented in Table 5. The feature configurations that generated the lowest errors with the validation subset were not those that generated the lowest errors

with the testing dataset. For $t0$, the KNN model applied to all 25 available variables generated the lowest errors. For the other time steps considered, it was the SVR models that generated the least errors. For time steps $t + 1$ and $t + 3$, the SVR models (FS4) using only local trend attributes generated the lowest errors. On the other hand, for time steps $t + 6$ and $t + 12$ the SVR (FS2) using local and spatially relevant trend attributes generated the lowest errors. The MAE and RMSE values for the testing dataset compare favorably with those generated by the validation subset for time steps $t0$, $t + 1$, and $t + 3$. However, the error values for the testing subsets relating to the $t + 6$ and $t + 12$ time steps are somewhat higher for the testing dataset compared to the validation subset.

Prediction errors associated with all 120 model evaluations relating to the testing subset are displayed in Figure 7. Whereas all the SML models evaluated for the $t0$ to $t + 3$ time steps generate errors for the testing subset within the ranges MAE 9–16 and RMSE 10–18, the models for the $t + 6$ and $t + 12$ timesteps generate more dispersed errors with MAE varying from 15 to 27 and RMSE varying from 19 to 30. For period $t + 12$, the XGB and KNN models all generate errors at the high end of those ranges, suggesting that the trained/validated XGB and KNN models for that timestep are to an extent overfitted. The XGB and KNN models are, therefore, not as generalizable for predicting independent datasets. On the other hand, the SVR models that generated relatively higher errors for the validation subsets for $t + 6$ and $t + 12$, compared to the other models (Figure 6), are much more generalizable when applied to datasets independent of the training/validation process.
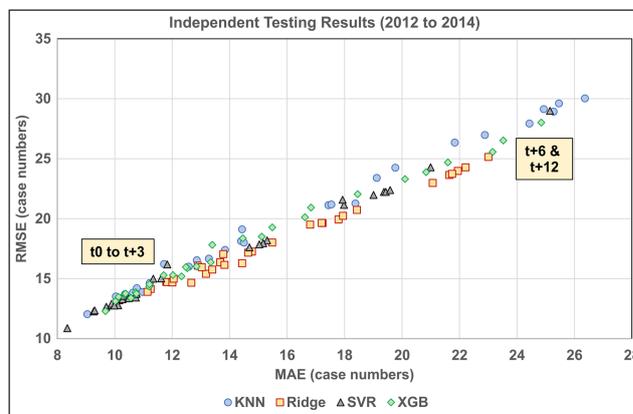
Figure 7. MAE versus RMSE for trained and validated SML models applied to the testing subsets (2012–2014) for time series $t0$, $t + 1$, $t + 3$, $t + 6$, and $t + 12$

4.4. Feature influences on high-performing test dataset forecasting models

The relative importance of each feature to the XGB models is readily revealed via tree-node selection coefficients. These are generated based on the node selections of the XGB model's underlying decision trees. These are displayed for selected high-performing models in Figure 8. For the KNN 25-variable $t0$ solution (Figure 8A), which generated the lowest prediction errors for the testing dataset for that time step, the features 5 (average for weeks $t − 1$ to $t − 3$), 0 (seasonality), 3 ($t − 2$ value), and 7 (average for weeks $t − 1$ to $t − 12$) exerted the greatest influence on the XGB forecast, in that order. These four attributes are all
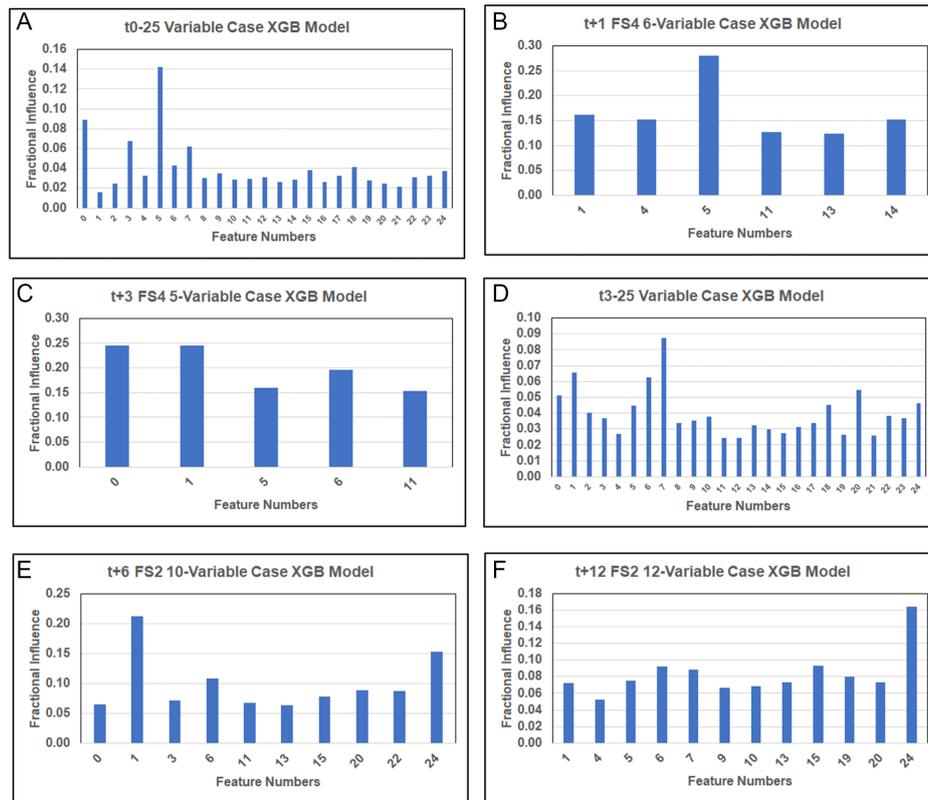
Figure 8. XGB feature influences on solutions applied to feature selections providing accurate case trend forecasts for specific time steps

derived from the local time series. For the $t+1$ FS4 solution (Figure 8B), involving six features derived from the local time series, feature 5 is again the most influential. For the $t+3$ FS4 solution (Figure 8C), involving just five features derived from the local time series, feature 5 is also included and all five features are assigned relatively high importance. However, for that XGB model, the two seasonality features, 0 and 1, are assigned the highest importance. The $t+3$ 25-variable model (Figure 8D), which performed well for that timestep, assigned features 7, 1, 6 (average for weeks $t-1$ to $t-6$), 20 (rate of change from weeks $t-11$ to $t-1$ for the BAC time series), and 0 the most importance.

For the $t+6$ FS2 XGB model (Figure 8E), features 1, 24 (rate of change from weeks $t-11$ to $t-1$ for the national time series), and 6 were assigned the most importance. For the $t+12$ FS2 XGB model (Figure 8F), features 24, 15 (rate of change from weeks $t-6$ to $t-1$ for the PES time series), 6, and 7 are assigned the most importance. Overall, considering all the time steps modeled by XGB, the locally derived weekly averages (features 5, 6, and 7) and seasonality (features 0 and 1) are the most influential for time steps $t0$, $t+1$, and $t+3$. On the other hand, for time periods $t+6$ and $t+12$, spatially derived rate of change information (features 15–24) is assigned equal or greater importance by the models than the locally derived weekly average and seasonality features.

4.5. Forecast comparisons with alternative univariate prediction methods

For the $t0$ time step of the Kom dataset, compared to the best trend-attribute model forecast for the testing dataset (MAE = 9.043; RMSE = 12.031), a naïve forecast applied to the univariate time

series generates slightly higher errors (MAE = 9.936; RMSE = 13.945), and a seasonally adjusted, 3-interval, MA model achieves almost the same error levels (MAE = 9.165; RMSE = 12.693). However, a standard 3-interval MA model applied to the univariate $t0$ time outperforms the trend-attribute model for the $t0$ time step (MAE = 7.945; RMSE = 11.240), as does an ARIMA(2,1,3) model (MAE = 7.853; RMSE = 11.216) and a univariate convolutional neural network CNN(1D) SDL model (MAE = 7.899; RMSE = 11.186). Clearly, these univariate models focusing only on a short prior time-interval provide a better forecasting alternative for the Komarom $t0$ dataset (Figure 9).

For time steps $t+1$, $t+3$, $t+6$, and $t+12$, the trend-attribute SML models substantially outperform an ARIMA(2,1,3) model applied to the univariate dataset (Figure 9). The SML models also outperform a multi-step encoder-decoder, bidirectional, long short-term memory (LSTM) model [34, 35] with a 3-period interval applied to the univariate datasets (Figure 9), by a relatively small amount for $t+6$, but more substantially for $t+1$, $t+3$, and $t+12$. The optimized bidirectional LSTM model applied involved the following configuration of structure and control parameters: one LSTM layer (bidirectional) with 200 nodes; no dropout; one dense layer with 100 nodes applying the "relu" activation function; one output layer with one node applying a linear activation function; "adam" optimizer; learning rate 0.0001; batch size 32; and 200 iterations run for time series $t0$, $t+1$, and $t+3$ and 750 iterations for time series $t+6$ and $t+12$.

The trend-attribute models offer the additional advantages over the univariate models of providing useful insight into the key influencing factors associated with the forecasts they generate. This latter information facilitates various data-mining applications
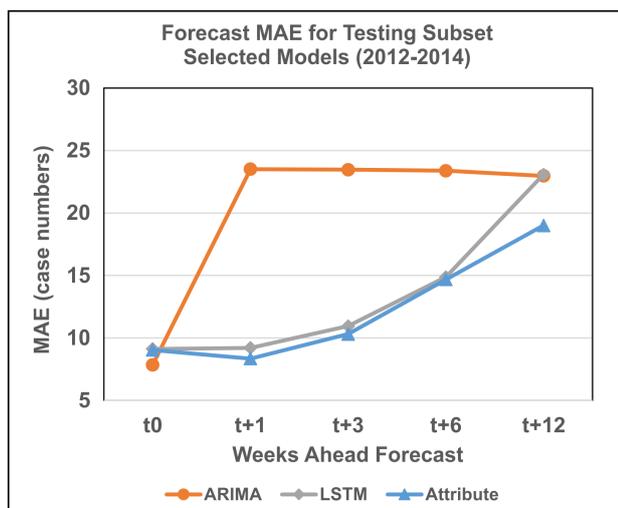
Figure 9. MAE forecasting error comparisons with other univariate, multi-step forecasting methods. RMSE displays similar relationships between the model forecasts

that the trend-attribute models can perform. Such data mining possibilities with case-number epidemiological time series are not possible with the univariate time-series models such as ARIMA, 1-D CNN, and LSTM models.

5. Discussion

The SML, trend-attribute SML models as developed and implemented for forecasting the Komarom VZV, weekly reported-case time-series provide a useful and effective forecasting method. They achieve this by incorporating historical information from the local univariate time series plus time-series data from other potentially influential regions. The SML-optimizer combinations applied to the training/validation subset make it possible to offer the SML models a substantial number of potentially influential attributes (up to 25 for the Komarom case). The optimizers then work to filter out the least influential of those attributes with respect to specific datasets. This approach is highly transparent and leads to solutions that generate meaningful forecasting performance from a limited set of optimizer-selected features (e.g., the SVR FS4 model with just five features [0 1 5 6 11] applied to the $t + 3$ time series, Figure 10). It also generates multiple feature combinations worthy of evaluating with specific datasets. Some of these feature combinations focus just on local information; others incorporate local plus regional information. Further consideration of the best-performing feature combinations with the XGB model can reveal which of the selected features are the most important to each high-performing feature combination selected.

The developed methodology could be readily applied to time-series datasets relating to other contagious diseases, particularly so where historical weekly data are available for multiple years but recorded separately from a number of regional locations within a specific country or geographical region. Further epidemiological case applications are required to evaluate the forecasting and data mining capabilities of the proposed method.

The spatio-temporal trend-attribute method is transparent and flexible in the sense that the attributes extracted from specific univariate time series are clearly identified. They are also easy to calculate and could be varied and/or expanded to suit the recorded trends. For instance, in this study only two attributes were extracted
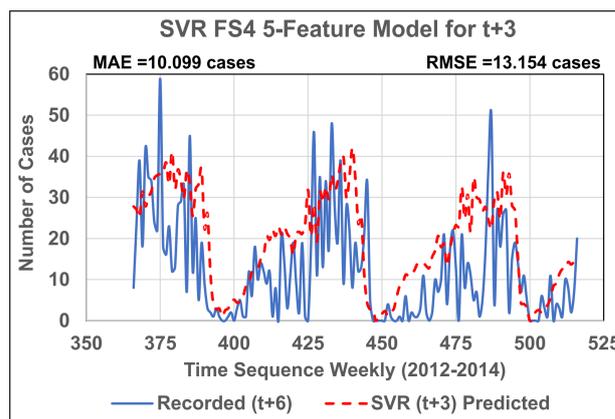
Figure 10. Actual versus predicted case numbers for the $t + 3$ test dataset applying the SVR FS4 (features [0 1 5 6 11]) model trained and validated with the 2005–2011 dataset

from the regional datasets available to provide a spatial influence dimension to the Komarom dataset studied (i.e., 5-week and 10-week rate of change in recorded case numbers). Clearly, rates of change in recorded case numbers over longer or shorter past time intervals could also be considered, together with the combinations of various weekly averages. An effective way to evaluate the likely influence of regional attributes on a specific location is to consider their R and p correlation coefficients. Future studies are planned to apply the model to other epidemiological time series.

6. Conclusions

Optimized and SML models are able to forecast the VZV cases likely to occur in the short term (up to 13 weeks forward) for regions of Hungary. This is achieved using a trend-attribute dataset covering the previous 12 weeks of historical weekly recorded cases. SML models are trained with historical weekly case data recorded over multiple years for specific regions. The trend attributes calculated from the regional time series include seasonality factors, multiple weekly averages, and rates of change in recorded VZV case numbers over the past 12 weeks. By considering the data available from multiple regions in the prediction of VZV cases in specific regions, both spatial and temporal dimensions are incorporated into the trend attributes considered. As well as being beneficial for weeks-ahead prediction purposes, the compiled trend-attribute dataset can be transparently data mined to reveal the historical influences on case numbers. The SML models are trained using weekly data from 2005 to 2011, and the trained models are then tested with weekly data from 2012 to 2014 for the dataset evaluated. The trend attributes considered are able to cope with the long-term decline in VZV case numbers over that period (due to vaccination programs) and seasonal fluctuations in case numbers. The SML model results presented focus specifically on the Komarom-Esztergom (Kom) county. However, the SML models consider the relative influences of other regional and national case trends on the Kom case trend, with those influential regions selected by optimization. Several optimizers are applied with the KNN SML model to establish those spatial and temporal trend-attribute features that exert the greatest influence on the Kom VZV case trends. The most accurate SML model forecasts for the Kom case trends are generated by trend-attribute models evaluated by SVR and XGB models. The predictions generated by these models for 1–4 weeks ahead ($t0$ to $t + 3$)

mostly exploit the trend-attribute data derived from the Kom time series itself. However, the most accurate SML forecasts for periods $t+6$ and $t+12$ exploit Kom time-series attributes together with some regional and national trend attributes. The SVR model proved to be the most generalizable of the SML models evaluated generating the lowest prediction errors.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in UC Irvine Machine Learning Repository at https://doi.org/10.24432/C5103B

References

[1] Brisson, M., Edmunds, W. J., Law, B., Gay, N. J., Walld, R., Brownell, M., . . . , & de Serres, G. (2001). Epidemiology of varicella zoster virus infection in Canada and the United Kingdom. *Epidemiology & Infection*, *127*(2), 305–314. https://doi.org/10.1017/S0950268801005921

[2] Amodio, E., Tramuto, F., Cracchiolo, M., Sciuto, V., de Donno, A., Guido, M., . . . , & Vitale, F. (2015). The impact of ten years of infant universal varicella vaccination in Sicily, Italy (2003–2012). *Human Vaccines & Immunotherapeutics*, *11*(1), 236–239. https://doi.org/10.4161/hv.36157

[3] Russell, M. L., Svenson, L. W., Yiannakoulias, N., Schopflocher, D. P., Virani, S. N., & Grimsrud, K. (2005). The changing epidemiology of chickenpox in Alberta. *Vaccine*, *23*(46–47), 5398–5403. https://doi.org/10.1016/j.vaccine.2005.05.008

[4] Sauboin, C., Holl, K., Bonanni, P., Gershon, A. A., Benninghoff, B., Carryn, S., . . . , & Wutzler, P. (2019). The impact of childhood varicella vaccination on the incidence of herpes zoster in the general population: Modelling the effect of exogenous and endogenous varicella-zoster virus immunity boosting. *BMC Infectious Diseases*, *19*, 126. https://doi.org/10.1186/s12879-019-3759-z

[5] Rafferty, E., McDonald, W., Qian, W., Osgood, N. D., & Doroshenko, A. (2018). Evaluation of the effect of chickenpox vaccination on shingles epidemiology using agent-based modeling. *PeerJ*, *6*, e5012. https://doi.org/10.7717/peerj.5012

[6] Wood, D. A. (2024). Weeks-ahead epidemiological predictions of varicella cases from univariate time series data applying artificial intelligence. *Infectious Diseases & Immunity, 4*(1), 25–34. https://doi.org/10.1097/ID9.0000000000000096

[7] Zhang, X., Liu, Y., Yang, M., Zhang, T., Young, A. A., & Li, X. (2013). Comparative study of four time series methods in forecasting typhoid fever incidence in China. *PLOS ONE*, *8*(5), e63116. https://doi.org/10.1371/journal.pone.0063116

[8] Abdulkareem, M., & Petersen, S. E. (2021). The promise of AI in detection, diagnosis, and epidemiology for combating COVID-19: Beyond the hype. *Frontiers in Artificial Intelligence*, *4*, 652669. https://doi.org/10.3389/frai.2021.652669

[9] Agrebi, S., & Larbi, A. (2020). Use of artificial intelligence in infectious diseases. In D. Barh (Ed.), *Artificial intelligence in precision health* (pp. 415–438). Academic Press. https://doi.org/10.1016/B978-0-12-817133-2.00018-5

[10] Absar, N., Uddin, N., Khandaker, M. U., & Ullah, H. (2022). The efficacy of deep learning based LSTM model in forecasting the outbreak of contagious diseases. *Infectious Disease Modelling*, *7*(1), 170–183. https://doi.org/10.1016/j.idm.2021.12.005

[11] Saleem, F., Al-Ghamdi, A. S. A. M., Alassafi, M. O., & AlGhamdi, S. A. (2022). Machine learning, deep learning, and mathematical models to analyze forecasting and epidemiology of COVID-19: A systematic literature review. *International Journal of Environmental Research and Public Health*, *19*(9), 5099. https://doi.org/10.3390/ijerph19095099

[12] Tat Dat, T., Frédéric, P., Hang, N. T., Jules, M., Duc Thang, N., Piffault, C., . . . , & Tien Zung, N. (2020). Epidemic dynamics via wavelet theory and machine learning with applications to Covid-19. *Biology*, *9*(12), 477. https://doi.org/10.3390/biology9120477

[13] Wang, Y., Xu, C., Wang, Z., Zhang, S., Zhu, Y., & Yuan, J. (2018). Time series modeling of pertussis incidence in China from 2004 to 2018 with a novel wavelet based SARIMA-NAR hybrid model. *PLOS ONE*, *13*(12), e0208404. https://doi.org/10.1371/journal.pone.0208404

[14] Wood, D. A. (2022). Near-term, national solar capacity factor forecasts aided by trend attributes and artificial intelligence. *International Journal of Energy and Environmental Engineering*, *13*(4), 1129–1146. https://doi.org/10.1007/s40095-022-00488-3

[15] Rozemberczki, B., Scherer, P., Kiss, O., Sarkar, R., & Ferenci, T. (2021). Chickenpox cases in Hungary: A benchmark dataset for spatiotemporal signal processing with graph neural networks. *arXiv Preprint: 2102.08100*. https://doi.org/10.48550/arXiv.2102.08100

[16] UC Irvine Machine Learning Repository. (2021). *Hungarian chickenpox cases*. https://doi.org/10.24432/C5103B

[17] Karsai, J., Csuma-Kovács, R., Dánielisz, Á., Molnár, Z., Dudás, J., Borsos, T., & Röst, G. (2020). Modeling the transmission dynamics of varicella in Hungary. *Journal of Mathematics in Industry*, *10*(1), 12. https://doi.org/10.1186/s13362-020-00079-z

[18] Statsmodels. (2023). *statsmodels.tsa.seasonal.seasonal_decompose*. Retrieved from: https://www.statsmodels.org/stable/generated/statsmodels.tsa.seasonal.seasonal_decompose.html

[19] Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review*, *57*(3), 238–247. https://doi.org/10.2307/1403797

[20] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

[21] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018

[22] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[23] Scikit-learn. (n.d.). *Machine learning in* Python. Retrieved from: https://scikit-learn.org/stable/

[24] Scikit-learn. (2011). *sklearn.model_selection.GridSearchCV*. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[25] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

[26] Wood, D. A. (2023). Ozone air concentration trend attributes assist hours-ahead forecasts from univariate recorded data avoiding exogenous data inputs. *Urban Climate*, *47*, 101382. https://doi.org/10.1016/j.uclim.2022.101382

[27] Scikit-learn. (2011). *Cross-validation: Evaluating estimator performance*. Retrieved from: https://scikit-learn.org/stable/modules/cross_validation.html

[28] Bilal, Pant, M., Zaheer, H., Garcia-Hernandez, L., & Abraham, A. (2020). Differential evolution: A review of more than two decades of research. *Engineering Applications of Artificial Intelligence*, *90*, 103479. https://doi.org/10.1016/j.engappai.2020.103479

[29] Joshi, A. S., Kulkarni, O., Kakandikar, G. M., & Nandedkar, V. M. (2017). Cuckoo search optimization: A review. *Materials Today: Proceedings*, *4*(8), 7262–7269. https://doi.org/10.1016/j.matpr.2017.07.055

[30] Rao, R. V. (2016). Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems. *International Journal of Industrial Engineering Computations*, *7*(1), 19–34. https://doi.org/10.5267/j.ijiec.2015.8.004

[31] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of International Conference on Neural Networks,* 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

[32] Abualigah, L., & Diabat, A. (2021). Advances in Sine Cosine Algorithm: A comprehensive survey. *Artificial Intelligence Review*, *54*(4), 2567–2608. https://doi.org/10.1007/s10462-020-09909-3

[33] Gini, C. (1997). Concentration and dependency ratios. *Journal of Economic Policy*, *87*, 769–790. https://www.mulino.it/subscriptions/issn/1120-9496

[34] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[35] Gu, J., Liang, L., Song, H., Kong, Y., Ma, R., Hou, Y., . . . , & Zhang, Y. (2019). A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in Guangxi, China. *Scientific Reports*, *9*(1), 17928. https://doi.org/10.1038/s41598-019-54495-2

Appendix A

A. Trend attributes calculated for other Hungarian regions considered likely to spatially influence the VZV case trends in the Komarom-Esztergom (Kom) region.

B. Historical trend attributes calculated and allocated to each data record in a time series. Trend attribute numbers 15–24 are those calculated for regions Bac, Bud, Gyo, and Pes (defined in Figure 1) and for the Hungarian national VZV case trend for use in forecasting the Kom cases for periods $t0$ to $t + 12$.

	Weekly VZV case trend attributes considered for regions spatially influencing Kom	Calculation
15	Rate of change from weeks $(t - 6)$ to $(t - 1)$	$[W(t - 6) - W(t - 1)]/5$ for PES
16	Rate of change from weeks $(t - 11)$ to $(t - 1)$	$[W(t - 11) - W(t - 1)]/10$ for PES
17	Rate of change from weeks $(t - 6)$ to $(t - 1)$	$[W(t - 6) - W(t - 1)]/5$ for GYO
18	Rate of change from weeks $(t - 11)$ to $(t - 1)$	$[W(t - 11) - W(t - 1)]/10$ for GYO
19	Rate of change from weeks $(t - 6)$ to $(t - 1)$	$[W(t - 6) - W(t - 1)]/5$ for BAC
20	Rate of change from weeks $(t - 11)$ to $(t - 1)$	$[W(t - 11) - W(t - 1)]/10$ for BAC
21	Rate of change from weeks $(t - 6)$ to $(t - 1)$	$[W(t - 6) - W(t - 1)]/5$ for BUD
22	Rate of change from weeks $(t - 11)$ to $(t - 1)$	$[W(t - 11) - W(t - 1)]/10$ for BUD
23	Rate of change from weeks $(t - 6)$ to $(t - 1)$	$[W(t - 6) - W(t - 1)]/5$ for entire nation
24	Rate of change from weeks $(t - 11)$ to $(t - 1)$	$[W(t - 11) - W(t - 1)]/10$ for entire nation

Note: $t - 1$ refers to 1 week before the current week ($t0$); $t - 12$ refers to 12 weeks before the current week ($t0$).