Lyme Disease Cases in the United States Projected Through 2012: Time Series Model Identification and Forecasts of the Federal CDC Data

By: Jeffrey Lidicker, MA and Laureen Leigh

Published by: WILDER Network, Inc.
World International Lyme Disease Emergency Rescue Network
www.wildernetwork.org

5/11/5

Key Words: Lyme Disease Cases, Lyme Disease Projection, Forecasting, Time Series Model Identification, AR(1), ARIMA(1, 1, 0), Datastat Services, WILDER Network

Contact Information:
Jeffrey Lidicker, MA
Datastat Services
PO Box 523
Pipersville, PA 18947-0523
215-813-8361
DatastatServices@AOL.com

If you or someone you know has Lyme disease data that needs to be analyzed or have an idea for a study and/or research you want to do or if you have Lyme disease and wish to participate in a study, please contact Jeffrey Lidicker.

INTRODUCTION

Lyme disease is one of a handful of tick-borne diseases available in North America. Originally discovered in Old Lyme Connecticut, this disease now has the distinction of being the most common of all tick borne diseases.[1] In fact, among all vector-borne diseases (includes, mosquitoes, leaches, etc.), Lyme disease is rapidly emerging in incidence.[1] The Federal Center for Disease Control (CDC) recognizes that their counts, being subject to strict guidelines, may underestimate the true number of cases by ten fold.[2] Of the 23,763 cases confirmed by the CDC in the United States for the year 2002 alone, it is estimated that 90% of them occur in the North Eastern most states.[3] As expenditures for health case in the United States continue to soar, the import of disease incidence rates only gets more serious. In order to better understand the societal and economic burden of the most common vector borne disease of Lyme disease, accurate predictions for the number of cases per year are warranted. The intent of this study is to predict up to 2012 the number of CDC reported Lyme disease cases per year in the US.

EXECUTIVE SUMMARY

The estimated number of CDC reported Lyme disease cases in the US in the year 2012 will be between 25,000 and 40,000 cases per year (see Table 10 and Figure 3) with a 95% level of confidence. In the minimum scenario, this represents a small increase from 2002 of only about 1,000 cases per year, but an increase of 15,900 per year in the maximum scenario (67% or 2/3 increase in ten years). However, a total of 32,000 cases per year are expected, producing 8,000 cases more in less than ten years (35% or 1/3 increase). This represents a steady increase in the number of cases per year and a subsequent additional burden on the health care resources. Considering that the CDC has suspected that their counts may under represent the actual number of Lyme disease cases by as much as ten fold, this prediction could actually mean as much as an additional 160,000 cases **per year** in ten years. The final Time Series estimation model performed well and is of type ARIMA(1, 1, 0) and represented by:

$$(1 + \mathbf{0.74131}*B)(1 - B)Z_t = \mathbf{1{,}065.9} + a_t$$

METHODS

The data for the study was garnished from the Federal CDC website were annual counts of cases of Lyme disease in the United States from 1982 through 2002.[4,5,6] The CDC only started recording these counts in 1982. Use of the standard Time Series ARIMA analysis is used to identify the model. Parameter estimation and forecasts are done using Maximum Likelihood Estimation. The software used for all statistical analysis is SAS[7].

RESULTS

The Data:
The original data has a clear upward trend and appears in Table 1 with the graph in Figure 1. In an effort to eliminate the trend, the first differences were calculated and appear in Figure 2. The differenced data appears to be stationary.

Table 1: Data on Lyme Disease Cases in the US from CDC

Year	Cases
1982	400
1983	600
1984	1,700
1985	2,900
1986	1,600
1987	2,700
1988	5,000
1989	8,800
1990	7,950
1991	9,470
1992	9,908
1993	8,257
1994	13,043
1995	11,700
1996	16,455
1997	12,801
1998	16,801
1999	16,273
2000	17,730
2001	17,029
2002	23,763

Figure 1: Lyme Disease Cases Data from CDC

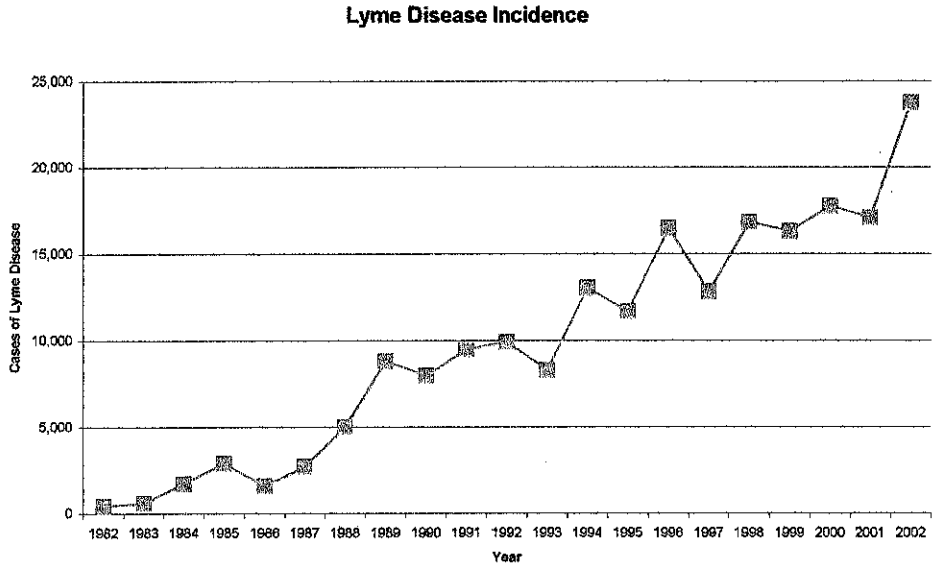
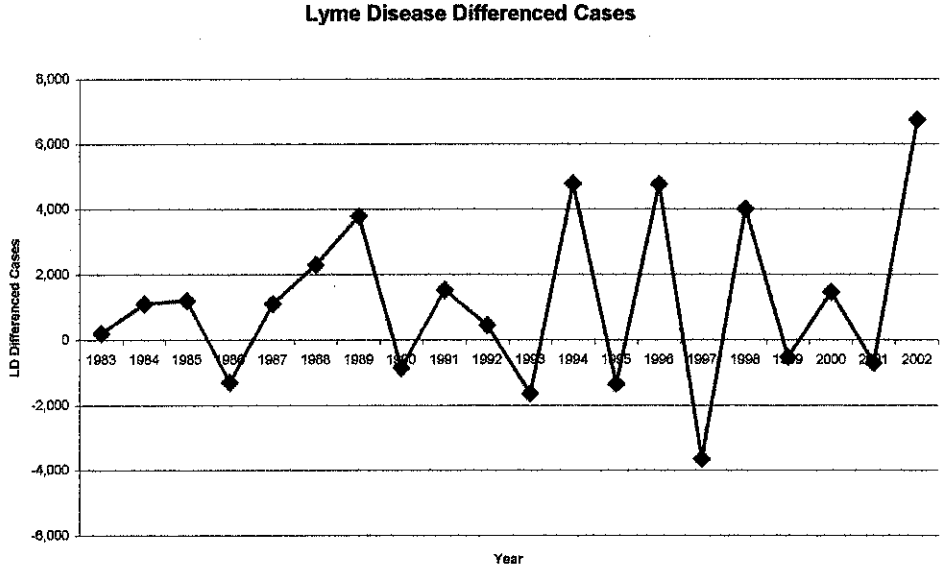


Figure 2: Lyme Disease Cases Data Differenced from CDC



Model Identification:
In order to determine the underlying model, an ACF and PACF analysis was performed on both the original data and the differenced data. The results for the original data appear in Table 2 and Table 3 while the results for the differenced data appear in Table 4 and Table5.

From the results of the analysis, it is clear the differencing is necessary as the ACF for the original data decays very slowly but then goes back up in magnitude again in a cyclic nature. The ACF for the differenced data decays slowly but does not come back. Since, the PACF for the differenced data cuts off after lag 1 (and the ACF decays slowly), the model is determined to be a classic example of an ARIMA(1,1,0).

Table 2: ACF Analysis for Original Lyme Disease Cases Data

The ARIMA Procedure
Name of Variable = z

Mean of Working Series	9756.19
Standard Deviation	6554.672
Number of Observations	21

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	42963720	1.00000	|********************|	0
1	32489660	0.75621	|*************** |	0.218218
2	29267108	0.68121	|************* |	0.319502
3	22676778	0.52781	|********** |	0.382460
4	19203340	0.44697	|********* |	0.415702
5	11793530	0.27450	|***** |	0.437989
6	8291439	0.19299	|**** |	0.446106
7	1561146	0.03634	|* |	0.450064
8	-1188819	-.02767	*| |	0.450204
9	-6069382	-.14127	***| |	0.450285
10	-7510301	-.17481	***| |	0.452390
11	-9903042	-.23050	*****| |	0.455595
12	-14143493	-.32920	*******| |	0.461115
13	-14750164	-.34332	*******| |	0.472174
14	-17855348	-.41559	********| |	0.483915
15	-17062628	-.39714	********| |	0.500622
16	-16853630	-.39228	********|	0.515406
17	-13743177	-.31988	******|	0.529433
18	-12097011	-.28156	******|	0.538557
19	-9347372	-.21756	****|	0.545522
20	-6240494	-.14525	***|	0.549638

Table 3: PACF Analysis for Original Lyme Disease Cases Data

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.75621	| . |********** . |
2	0.25540	| . |***** . |
3	-0.12217	| . **| . |
4	0.01229	| . | . |
5	-0.20777	| . ****| . |
6	-0.01248	| . | . |
7	-0.15644	| . ***| . |
8	-0.00532	| . | . |
9	-0.07593	| . **| . |
10	-0.00166	| . | . |
11	0.00090	| . | . |
12	-0.27682	| . ******| . |
13	0.06606	| . |* . |
14	-0.20608	| . ****| . |
15	0.07120	| . |* . |
16	0.02212	| . | . |
17	0.05645	| . |* . |
18	0.06874	| . |* . |
19	-0.13044	| . ***| . |
20	0.15821	| . |*** . |

Table 4: ACF Analysis for Differenced Lyme Disease Cases Data

The ARIMA Procedure
Name of Variable = z

Mean of Working Series	1168.15
Standard Deviation	2538.602
Number of Observations	20

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	6444502	1.00000	|||||||||||||||||||||	0
1	-3832235	-.59465	||||||||||||| . |	0.223607
2	2487089	0.38592	| . |||||||| . |	0.292166
3	-2676739	-.41535	| . |||||||| . |	0.316630
4	1994492	0.30949	| . |||||| . |	0.342792
5	-1718270	-.26663	| . ||||| . |	0.356490
6	828100	0.12850	| . |||| . |	0.366325
7	281466	0.04368	| . |* . |	0.368571
8	55634.884	0.00863	| . | . |	0.368830
9	-91936.793	-.01427	| . | . |	0.368840
10	-634236	-.09842	| . |** . |	0.368868
11	653127	0.10135	| . |** . |	0.370178
12	-1033290	-.16034	| . |*** . |	0.371563
13	682288	0.10587	| . |** . |	0.375007
14	506786	0.07864	| . |** . |	0.376498
15	80858.341	0.01255	| . | . |	0.377318
16	-608722	-.09446	| . |** . |	0.377339
17	1250.238	0.00019	| . | . |	0.378520
18	71515.245	0.01110	| . | . |	0.378520
19	-269429	-.04181	| . |* . |	0.378536

Table 5: PACF Analysis for Differenced Lyme Disease Cases Data

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.59465	| ********** | . . |
2	0.04999	| . . |* . . |
3	-0.25995	| . *****| . . |
4	-0.06147	| . . *| . . |
5	-0.09468	| . . **| . . |
6	-0.20550	| . . ****| . . |
7	0.16332	| . . |*** . . |
8	0.07903	| . . |** . . |
9	-0.01314	| . . | . . |
10	-0.07644	| . . **| . . |
11	-0.01854	| . . | . . |
12	-0.12038	| . . **| . . |
13	-0.10689	| . . **| . . |
14	0.21304	| . . |**** . . |
15	0.09343	| . . |** . . |
16	-0.08529	| . . **| . . |
17	-0.01128	| . . | . . |
18	-0.02325	| . . | . . |
19	-0.02932	| . . *| . . |

Parameter Estimation:
In order to estimate the parameters for the ARIMA(1,1,0) model, I used the Maximum Likelihood option within SAS. In this case, *phi* 1 = -0.74131 ($p = 0.0001$) from the ACF at lag 1 on the differenced data. From the differenced data, *mu* = 1,065.9 ($p < 0.0001$). The Sample Standard Error = 2,605 and the *sigma-a-hat squared* = 3,774,419. These results are summarized in Table 6.

So, the final model is ARIMA(1,1,0) of the form:

$$(1 + \mathbf{0.74131}*B)(1 - B)Z_t = \mathbf{1{,}065.9} + a_t$$

Table 6: Lyme Disease SAS Output for Parameter Estimation Procedure

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	1065.9	250.59867	4.25	<.0001	0
AR1,1	-0.74131	0.19350	-3.83	0.0001	1

Constant Estimate	1856.128
Variance Estimate	3774419
Std Error Estimate	1942.786
AIC	362.323

Model Performance:
To evaluate the model validity, the residuals from the model were evaluated. The ACF and PACF of the residuals show no significant spikes or patterns as can be seen in Tables 7 and 8. The Q-statistic for white noise shows no significance for rejecting the hypothesis that the residuals are white noise ($p = 0.69$). See Table 9 for SAS output of Q-statistics testing for white noise. Therefore, we conclude that our model performs just fine.

Table 7: Model Residuals ACF Results

The ARIMA Procedure
Name of Variable = RESIDUAL

Mean of Working Series	-32.1036
Standard Deviation	1848.391
Number of Observations	20

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	3416550	1.00000	|********************|	0
1	-255816	-.07488	| . *| . |	0.223607
2	-685684	-.20069	| . ****| . |	0.224857
3	-664096	-.19438	| . ****| . |	0.233642
4	-271959	-.07960	| . **| . |	0.241592
5	-423878	-.12407	| . **| . |	0.242900
6	657406	0.19242	| . |**** . |	0.246048
7	685033	0.20050	| . |**** . |	0.253460
8	141952	0.04155	| . |* . |	0.261270
9	-944923	-.27657	| . ******| . |	0.261601
10	117164	0.03429	| . |* . |	0.275833
11	-117123	-.03428	| . *| . |	0.276047
12	-366043	-.10714	| . **| . |	0.276259
13	680663	0.19923	| . |**** . |	0.278329
14	592788	0.17350	| . |*** . |	0.285370
15	-211753	-.06198	| . *| . |	0.290597

Table 8: Model Residuals PACF Results

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.07488	| . *| . |
2	-0.20746	| . ****| . |
3	-0.23966	| . *****| . |
4	-0.19039	| . ****| . |
5	-0.30029	| . ******| . |
6	-0.00684	| . | . |
7	0.08435	| . |** . |
8	0.06624	| . |* . |
9	-0.19721	| . ****| . |
10	0.06557	| . |* . |
11	-0.02438	| . | . |
12	-0.18653	| . ****| . |
13	0.10399	| . |** . |
14	0.06948	| . |* . |
15	0.04956	| . |* . |

Table 9: Q-statistic Test Results from SAS for Model Residual White Noise

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	3.88	6	0.6927	-0.075	-0.201	-0.194	-0.080	-0.124	0.192
12	9.11	12	0.6939	0.201	0.042	-0.277	0.034	-0.034	-0.107

Forecasts:

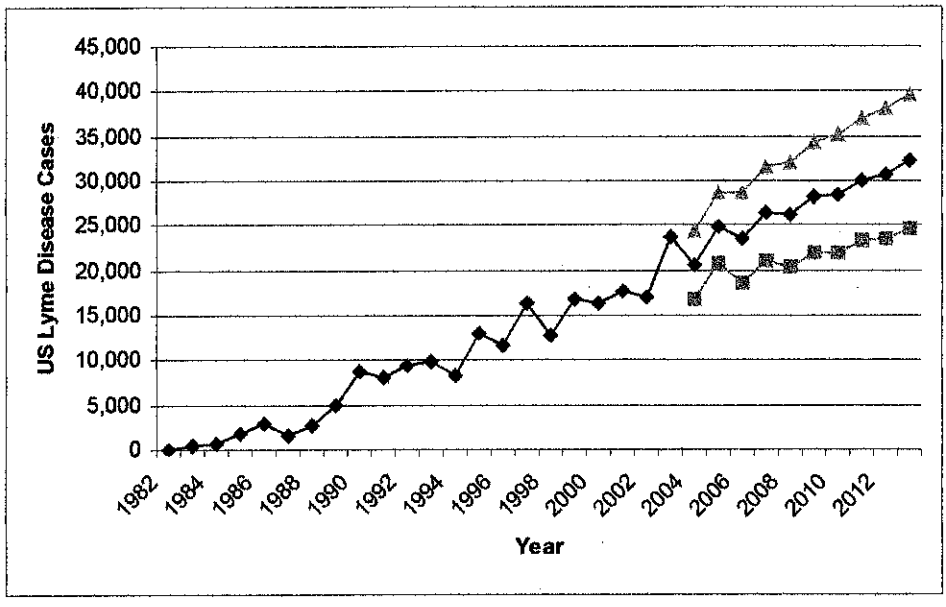
After calculating the predictions for the next ten years and their associated forecast limits, we get the following results summarized in Table 10 and illustrated in Figure 3.

Table 10: US Lyme Disease Predictions and Their 95% Forecast Limits

Forecasts for variable z

Year	Forecast	Std Error	95% Confidence Limits	
2003	20627.12	1942.79	16819.33	24434.91
2004	24807.92	2006.74	20874.78	28741.05
2005	23564.77	2548.06	18570.67	28558.87
2006	26342.46	2664.41	21120.32	31564.60
2007	26139.45	2993.93	20271.46	32007.44
2008	28146.07	3135.20	22001.19	34290.96
2009	28514.66	3376.30	21897.23	35132.10
2010	30097.55	3525.27	23188.15	37006.95
2011	30780.26	3721.06	23487.11	38073.41
2012	32130.29	3869.04	24547.11	39713.46

Figure 3: Cases of CDC Lyme Disease in the US with Forecasts and 95% Forecast Limits



CONCLUSIONS

The estimated number of CDC reported Lyme disease cases in the US in the year 2012 will be between 25,000 and 40,000 cases per year (see Table 10 and Figure 3) with a 95% level of confidence. In the minimum scenario, this represents a small increase from 2002 of only about 1,000 cases per year, but an increase of 15,900 per year in the maximum scenario (67% or 2/3 increase in ten years). However, a total of 32,000 cases per year are expected, producing 8,000 cases more in less than ten years (35% or 1/3 increase). This represents a steady increase in the number of cases per year and a subsequent additional burden on the health care resources. Considering that the CDC has suspected that their counts may under represent the actual number of Lyme disease cases by as much as ten fold, this prediction could actually mean as much as an additional 160,000 cases **per year** in ten years. The final Time Series estimation model performed well and is of type ARIMA(1, 1, 0) and represented by:

$$(1 + \mathbf{0.74131}*B)(1 - B)Z_t = \mathbf{1,065.9} + a_t$$

Future studies which do similar predictions by individual state would produce information that would allow health officials to concentrate resources in areas that show the strongest and fastest increases or largest numbers. States with possibly hostile climates for ticks (such as Arizona) should show slower growth rates than other states. However, this remains to be proven. In any event, a by state analysis will help Lyme disease epidemiologists better understand factors effecting tick populations. Further, as states themselves enact new legislation on reporting guidelines for this disease of increasing import, relative changes in infection rates as a function of these new policies may be detected. As the reporting guidelines are a science in of themselves, a full analysis over time of relative infection rates by state along with a review of the various state reporting legislation may help to optimize these issues. Certainly, if all states adopted similar and effective reporting guidelines sooner than later, a true more reliable understanding of the magnitude of the Lyme disease problem and its idiosyncrasies can be better monitored and eventually managed.

Currently, the WILDER Network is seeking funding for this proposed study if there are any readers who know of any resources that could be interested. As the Lyme disease community is a relatively small group, working together is important and we appreciate all forms of help.

REFERENCES

1. Center for Disease Control and Prevention. Lyme Disease Epidemiology. Division of Vector-Borne Infectious Diseases 2005. http://www.cdc.gov/ncidod/dvbid/lyme/epi.htm

2. Lyme Disease in the US, National Lyme Disease Data 1982-2000 http://www.stopticks.org/epidemiology/lymeus.asp

3. Orloski KA, Hayes EB, Campbell GL, and Dennis DT. Surveillance for Lyme Disease --- United States, 1992 – 1998. National Center for Infectious Diseases, Division of Vector-Borne Infectious Diseases 2000; 49(SS03); 1 – 11. http://cdc.gov/epo/mmwr/preview/mmwrhtml/ss490311.htm

4. Center for Disease Control and Prevention. Lyme Disease – United States, 1996. MMWR June 13, 1997; 46(23); 531-535. http://www.cdc.gov/mmwr/preview/mmwrhtml/00047906.htm

5. Center for Disease Control and Prevention. Lyme Disease Cases Reported to CDC by State Health Departments 1990-2001. http://www.cdc.gov/ncidod/dvbid/lyme/ldcases90-2001.htm,

6. Center for Disease Control and Prevention. Notice to Readers: Final 2002 Reports of Notifiable Diseases. MMWR. 8 August 2003; 52(31): 741-750. http://www.cdc.gov/mmwr.preview/mmwrhtml/mm5231a7.htm

7. SAS Institute Inc., Cary NC 27513