NBSIR 73-138

Time Series Forecasting of Highway Accident Fatalities

Alexander R. Craw

Institute for Applied Technology
Technical Analysis Division
National Bureau of Standards
Washington, D. C. 20234

March 1973

Final Report

Prepared for
National Highway Traffic Safety Administration
Department of Transportation
Washington, D. C. 20590

NBSIR 73-138

TIME SERIES FORECASTING OF HIGHWAY ACCIDENT FATALITIES

Alexander R. Craw

Institute for Applied Technology
Technical Analysis Division
National Bureau of Standards
Washington, D. C. 20234

March 1973

Final Report

Prepared for
National Highway Traffic Safety Administration
Department of Transportation
Washington, D. C. 20590

U. S. DEPARTMENT OF COMMERCE, Frederick B. Dent, Secretary

NATIONAL BUREAU OF STANDARDS, Richard W. Roberts, Director

ACKNOWLEDGEMENT

Particular thanks are due to Mrs. E. Leyendecker, who ran all of the computer programs and carried most of the computational analysis load. Also, to Dr. Julius Lieblein for many helpful suggestions and for discussions that provided insight into what had often been obscure.

TABLE OF CONTENTS

SUMMARY

Using twelve years of time series data on highway fatalities, the methodology currently employed by the National Highway Traffic Safety Administration (NHTSA) to forecast the annual (calendar year) total of highway accident fatalities were compared with those obtained by several computer routines based on exponential smoothing techniques and available at the National Bureau of Standards. The use of unadjusted and seasonally adjusted data was also examined.

It is found that there is no coercive evidence to lead to abandoning the present NHTSA methods in favor of readily available computer routines based on exponential smoothing methods.

Of the methods examined in this study, the best results were obtained with the EXPSMOOTHING routine using unadjusted fatality data.

1. The Problem: The Technical Analysis Division has been requested to examine the National Highway Traffic Safety Administration's (NHTSA) current methodologies for forecasting the annual total of highway accident fatalities given the monthly fatalities respectively for the first three months of the year, the first six months and finally the first nine months of the year, make comparisons of the results of their methodologies with those based on readily available computer programs at the National Bureau of Standards for time series analysis and forecasting.

2. The Data and Auxiliary Information Available: Data for each month are available for the twelve (12) year period 1960 through 1971. Also available are the outputs of NHTSA's Computer Program TIMSR4: Time Series Analysis of Fatalities and Vehicle-mileage Data. This time series analysis program provides 12-month moving averages, monthly seasonal indices, seasonally adjusted data for each month and the ordered differences between the seasonally adjusted data and an uncentered twelve month moving average. The ordered differences are used to determine for any month of a time series upper and lower bounds for the trend point such that the seasonally adjusted value will fall within those limits with a known frequency.

The TIMSR4 Program does not have an automatic extrapolation (forecasting ahead) feature. The NHTSA Methods of forecasting that we were asked to review are discussed in Section 5.

3. <u>Readily Available Computer Programs:</u> A constraint imposed on the study was to use computer programs that were readily available at the National Bureau of Standards. Thus, a comprehensive study of methods is not given here.

Several time-sharing interactive systems were immediately available. These were DIALCOM, Computer Sciences Corporation Conversational Executive and General Electric Mark II. Of these time-sharing programs we chose to use the General Electric EXPSMOOTHING routine, because of the variety of options available and the existence of an automatic extrapolation feature.

The Large Scale Systems STAT-PACK for the UNIVAC 1108 lists several routines for time series analysis. These are UNIVAC system subroutines and are not necessarily National Bureau of Standards routines, but they are available for use on the NBS UNIVAC 1108. The most appropriate subroutine for the problem at hand appeared to be GEXSMO - Generalized Exponential Smoothing.

Appendix B contains a discussion of the available computer programs at NBS as well as a discussion of the background and details of the GEXSMO computer program. Appendices A and C provide the same information for the EXPSMOOTHING program.

4. <u>Plan of Analysis:</u>

a. Use the General Electric - EXPSMOOTHING routine and the UNIVAC STAT-PACK GEXSMO routine.

b. Use both seasonally adjusted data and seasonal factors (as determined by the TIMSR4 program) and unadjusted data in separate runs.

c. Base forecasts on 4, 6, 8, and 11 years of data plus data for 3, 6, and 9 months of new data.

d. Do c. for each of the NHTSA forecasting algorithms.

5. Production Runs: We started production runs using the General Electric Time-share routine EXPSMOOTHING, using seasonally adjusted data that was determined by NHTSA's TIMSR4 computer program. We always chose the most complete output option, so that the routine always performed a cyclic analysis: if a cycle was indicated by the routine, the analysis part of the program took this into account. (We view this as a not too interesting bonus of the routine.) Horizon time was always chosen to exceed the lead time.* The statistical analysis of the output was accomplished by a mixture of computer and hand calculations. The results of this group of calculations are contained in Table 1.

The total amount of data available for forecasting covered 11 years, with a minimum of 3 years of data considered

---

*In the EXPSMOOTHING Program, the lead time is the term used to describe the time (in periods) between the last period of data one has as data as input and the period one is most interested in estimating correctly. The forecast will be optimized according to the lead time that is entered. The forecast horizon indicates how many periods beyond the last data point that will be forecasted, but has nothing to do with the optimization. (See Appendix A).

necessary for suitable application of the exponential smoothing process. Thus, annual "forecasts" could be made after 3, 6, and 9 months in each of 8 years. Instead of analyzing this full set of 24* periods, half of these were chosen, namely the three periods in those years with 4, 6, 8, and 11 antecedent years of data.

In the headings of Table 1, and subsequent tables, the following symbols were adopted:

S = the actual value of the time series datum (in this case, fatalities)

$\hat{S}$ = an estimate of S

$\tilde{S}$ = seasonally adjusted value of S

$\Delta = \hat{S} - S$; if $\left\{\Delta \begin{smallmatrix} < \\ > \end{smallmatrix} 0\right\}$ the estimate is $\left\{\begin{matrix} \text{below} \\ \text{above} \end{matrix}\right\}$ the actual value

Σ = summation over the lead time

Σ^* = summation over the calendar year (Jan. - Dec.)

$\Sigma\Delta/\Sigma S$ = (signed) fractional error over lead time

$\Sigma^*\Delta/\Sigma^*$ = (signed) fractional error over the calendar year (Jan. - Dec.)

MAD = mean absolute deviation

α = smoothing constant

k = length of cycle; $k = 1$ indicated data is not cyclic.

One of the interesting features of Table 1 is that there is no apparent pattern for the cycles with the age of the data. The same holds true for the α value (smoothing constant) and the order of smoothing. This led us to dupli-

---

*An additional 3 periods were analyzed for non-calendar twelve-month years, raising the potential sample size to 27 and the number tested to 15.

cate these computations by suppressing the cyclic analysis. The summary of the results of these computer runs is given in Table 2. Here there does appear to be some reasonableness in the variation of the smoothing constant with data age. The comparison of the forecasting results (Table 3) shows a slight advantage to the case of using the cyclic analysis, but the results are not heavily in favor of the latter.

In the case of 4 years and 3 months of seasonally adjusted data, the computer routine selected $\alpha = 0.175$ (order 2) for the optimum smoothing constant. We were interested in determing whether or not the minimum was sharp or flat, so we made runs for $\alpha = 0.170$ and 0.180. The order of smoothing remained at 2 and the minimum was found to be flat; only a very slight change in forecast error fraction could be gained by changing α from .175 to .170 (see Table 4). The change is from 0.0121 to 0.0109.

Table 5 presents the results of duplicate runs using unadjusted data in the EXPSMOOTHING routine. Cyclic analysis was permitted and each time the routine selected K=12 for the dominant cycle.

We also made matching runs starting with unadjusted data on the UNIVAC 1108 program GEXSMO, starting with a given cycle (seasonality) length of 12 and using 2 initial periods to determine the initial seasonality factors and other starting values. The GEXSMO routine always produces a forecast for the time indicated as the length of the cycle starting from the last input datum. In the language of the EXPSMOOTHING routine, the lead time is 1 and the horizon time is the length of the cycle. The summary statistics for these runs are presented in Table 6.

It is interesting to note that in working with unadjusted data and these two computer routines, it is possible to compare the seasonally adjusted data and seasonality factors produced by these routines with those determined by the TIMSR4 program. We have not made these comparisons, preferring at this stage to look at the end results.

Table 7 presents the detailed calculations for the current methodologies employed by NHTSA to forecast the calendar year fatalities. These methods were developed in accordance with descriptions furnished by Mr. Donald F. Mela in his letter of November 30, 1972, and have been labelled 1, 2 and 3 in the same order as described by Mr. Mela. Higher order methods are possible by taking into account more years of input data in the averaging process, but we have limited ourselves to the first three methods.

The notations used in the column headings of Table 7 are:

m_o = current total for specified months of fatalities

m_{-i} = the corresponding total for specified months of fatalities for i^{th} previous year.

Y_{-i} = the total fatalities for the i^{th} previous year

λ= leap year adjustment factor

$\hat{Y}_j$ = the forecast calendar year fatalities by method j

With this notation the NHTSA forecast methods may be written:

Method 1: $\hat{Y}_1 = (m_o/m_{-1})\, Y_{-1}$,

Method 2: $\hat{Y}_2 = \lambda \hat{Y}_1$

Method 3: $\hat{Y}_3 = \lambda (m_o/\sum_1^3 m_{-i}) \sum_1^3 Y_{-i}$

Method k: $\hat{Y}_k = \lambda (m_o/\sum_1^k m_{-i}) \sum_1^k Y_{-i}$, $k > 1$

Table 8 presents a composite summary of the fractional errors by each of the methods considered. In this table and Table 9, SAC = Seasonally adjusted data with cyclic analysis.

Considering just the NHTSA's methods, we find a gradual improvement in the forecast with increasing order of the method. However, in certain specific cases Method 2 gives slightly better results with methods of order greater than 2. Generally, the error decreases with the shorter lead time.

Comparing the four exponential smoothing methods, the sample evidence seems to favor EXPSMOOTHING using unadjusted data. However, each of the four methods has a mean absolute deviation of approximately 1 percent, which is virtually indistinguishable from that of Method 3.

Comparing the "best" NHTSA method (Method 3) with the "best" exponential smoothing routine we find for the sample of 12 trials out of the possible 24 forecasts that method 3 was better 5 times and the best exponential smoothing method was better on 7 occasions. This is obtained by comparing the entries in columns headed $\Sigma^*\Delta/\Sigma^*S$ and Unadj C in Table 8. However, the range of the absolute deviations of the exponential methods is smaller than that of any of the methods 1, 2 or 3. Thus, it would appear that there is a small gain in going to the exponential smoothing methods.

Our experience in using the EXPSMOOTHING routine has made us wary of using long lead times, in particular L = 12. In some cases we have found that better results can be obtained in making forecasts for a year ahead by using L = 3 in place of 12. See Table 9 for the appropriate comparisons. However, no clear cut rule can be recommended.

6. Conclusions

For the forecasting problem at hand (see section 1), if one can live with the percentage range of errors indicated with the present NHTSA's methods, there is no coercive evidence to lead to abandoning the present methods in favor of readily available computer routines for exponential smoothing methods.

Of the methods examined in this study, the best results were obtained with the EXPSMOOTHING routine using unadjusted fatality data.

Table 1. Error Analysis of Forecasts of Fatalities Using EXPSMOOTHING (General Electric) and Seasonally Adjusted (TIMSR4) Data for Calendar Year: January-December

Cyclic Analysis Permitted

Years of Forecast	Length of Record Years, Months	Cyclic	K	α	Order	MAD	Lead Time	$\Sigma\Delta$	$\Sigma^*\Delta$	ΣS	Σ^*S	$\Sigma\Delta/\Sigma S$	$\Sigma^*\Delta/\Sigma^*S$	$\Sigma\tilde{\Delta}$	$\Sigma\tilde{S}$	$\Sigma\tilde{\Delta}/\Sigma\tilde{S}$	Max $\tilde{\Delta}$
1964	4,0	C	4	.13462	2	123.48	12	-2176	-2176	47700	47770	-.0456	-.0456	-2425	47740	-.0508	193
	4,3	N	1	.17500	2	122.23	9	579	579	37720	47770	.0154	.0121	457	35564	.0129	470
	4,6	N	1	.17857	2	120.53	6	985	985	26490	47770	.0372	.0206	905	23915	.0378	409
	4,9	N	1	.18750	2	117.30	3	609	609	13280	47700	.0459	.0128	432	11900	.0363	289
1966	6,0	C	17	.13462	2	152.63	12	-1271	-1271	53041	53041	-.0240	-.0240	-1262	53043	-.0238	99
	6,3	C	17	.17500	2	112.04	9	-1000	-1000	42284	53041	-.0237	-.0189	-1115	40036	-.0279	235
	6,6	C	17	.17857	2	118.42	6	863	863	29284	53041	.0295	.0163	760	26510	.0287	323
	6,9	C	17	.37500	1	109.11	3	393	393	14597	53041	.0269	.0074	354	13134	.0270	171
1968	8,0	C	23	.13462	2	104.02	12	770	770	54862	54862	.0140	.0140	798	54919	.0145	614
	8,3	C	17	.17500	2	141.73	9	-635	-635	43518	54862	-.0146	-.0116	-650	41210	-.0158	363
	8,6	C	17	.25000	1	118.34	6	238	238	30075	54862	.0079	.0043	162	27297	.0060	389
	8,9	C	17	.37500	1	103.48	3	572	572	15153	54862	.0377	.0104	479	13707	.0349	418
1971	11,3	C	20	.12500	2	133.71	9	-939	-939	43870	55000	-.0214	-.0171	-810	41486	-.0195	256
	11,6	C	20	.25000	1	119.50	6	-520	-520	30290	55000	-.0172	-.0095	-516	27568	-.0187	114
	11,9	C	20	.37500	1	114.52	3	132	132	15010	55000	.0088	.0024	102	13588	.0075	158

Note: In this table the values of $\sum\Delta$ and $\sum^*\Delta$ are the same. For example, for the case 4,9 $\sum\Delta$ is the error over the lead time (October, November, December); since the error for each of the months January through September is exactly zero (they are known exactly) we have $\sum\Delta \equiv \sum^*\Delta$. In some tables $\sum\Delta \neq \sum^*\Delta$ in all cases (see Table 4 for example).

Table 2. Forecasts of Fatalities Using EXPSMOOTHING (General Electric) and Seasonally Adjusted (TIMSR4) Data for Calendar Year. January-December

Cyclic Analysis Omitted

Years of Forecast	Length of Record Years, Month	Cyclic	K	α		MAD	Lead Time	$\Sigma\Delta$	ΣS	Σ^*S	$\Sigma\Delta/\Sigma S$	$\Sigma^*\Delta/\Sigma^*S$	$\Sigma\tilde{\Delta}$	$\Sigma\tilde{S}$	$\Sigma\tilde{\Delta}/\Sigma\tilde{S}$	Max. $\tilde{\Delta}$
1964	4,0	N	1	.13462	2	130.97	12	-2210	47700	47700	-.0463	-.0463	-2455	47740	-.0514	784
	4,3	N	1	.17500	2	122.23	9	579	37720	47770	.0154	.0121	457	35564	.0129	470
	4,6	N	1	.17857	2	120.53	6	985	26490	47770	.0372	.0206	905	23915	.0378	409
	4,9	N	1	.18750	2	117.30	3	609	13280	47700	.0459	.0128	432	11900	.0363	289
1966	6,0	N	1	.13462	2	137.43	12	-1761	53041	53041	-.0332	-.0332	-1727	53043	-.0326	517
	6,3	N	1	.15	2	125.26	9	-1060	42284	53041	-.0251	-.0200	-1154	40036	-.0288	619
	6,6	N	1	.14286	2	127.25	6	913	29284	53041	.0312	.0172	795	26510	.0300	336
	6,9	N	1	.125	2	117.92	3	650	14597	53041	.0446	.0123	585	13134	.0445	249
1968	8,0	N	1	.09615	2	148.32	12	601	54862	54862	.0110	.0110	596	54919	.0109	589
	8,3	N	1	.125	2	140.58	9	-479	43518	54862	-.0110	-.0087	-497	41210	-.0121	405
	8,6	N	1	.10714	2	123.90	6	516	30075	54862	.0172	.0094	423	27297	.0155	345
	8,9	N	1	.125	2	123.48	3	255	15153	54862	.0168	.0047	193	13707	.0141	352
1971	11,3	N	1	.125	2	140.56	9	-783	43870	55000	-.0178	-.0142	-688	41486	-.0166	405
	11,6	N	1	.25	1	133.24	6	-327	30290	55000	-.0108	-.0060	-340	27568	-.0123	474
	11,9	N	1	.25	1	128.22	3	323	15010	55000	.0215	.0059	281	13588	.0207	319

Note: In this table $\sum\Delta = \sum^*\Delta$ and the column $\sum^*\Delta$ is omitted.

Table 3: Comparison of Results from EXPSMOOTHING

with and without Use of Cyclic Analysis for Seasonally Adjusted Data (TIMSR4)

Year of Forecast	Length of Record Years/Months	∑*Δ/∑*S (Cumulative Fractional Error)								Lower Error	
		L = 12		L=9		L=6		L=3		*	
		C	N	C	N	C	N	C	N	C	N
1964	4,0 4,3 4,6 4,9	-.0456*	-.0463	.0121^{T}	.0121^{T}	.0206^{T}	.0206^{T}	.0128^{T}	.0128^{T}	1 T T T	T T T
1966	6,0 6,3 6,6 6,9	-.0240*	-.0332	-.0189*	-.0200	.0163*	.0172	.0074*	.0123	1 1 1 1	
1968	8,0 8,3 8,6 8,9	.0140	.0110*	-.0116	-.0087*	.0043*	.0094	.0104	.0047*	1	1 1 1
1971	11,3 11,6 11,9	--	--	-.0171	-.0142*	-.0095	-.0060*	.0024*	.0059	1	1 1
	*	2	1	1	2	2	1	2	1	7	5
	Ties			1T	1T	1T	1T	1T	1T	3	3

Table 4: Miscellaneous Comparisons Using EXPSMOOTHING and Seasonally Adjusted Data (TIMSR4) for Calendar Year:

January - December

Year of Forecast	Length of Record Years, Months	Cyclic	K	α	Order	MAD	Lead Time	$\sum\Delta$	$\sum{*}\Delta$	$\sum S$	$\sum{*}S$	$\sum\Delta/\sum S$	$\sum{*}\Delta/\sum{*}S$
1964	4,3	N	1	.17500	2	122.23	9	579	579	37720	47700	.0154	.0121
	4,3	N	1	.17000	2	122.15	9	521	521	37720	47700	.0138	.0109
	4,3	N	1	.18000	2	122.53	9	635	635	37720	47700	.0168	.0133
1964	4,0	C	4	.18750	2	109.60	3	-460	-1212	9980	47700	-.0461	-.0254
	4,0	C	4	.13462	2	123.48	12	-2176	-2176	47700	47700	-.0456	-.0456
1966	6,0	C	17	.25000	2	101.43	3	119	2517	10757	53041	.0110	.0457
	6,0	C	17	.13462	2	152.63	12	-1271	-1271	53041	53041	-.0240	-.0240
1968	8,0	C	23	.12500	3	87.43	3	- 113	-1820	11344	54862	-.0100	-.0332
	8,0	C	23	.13462	2	104.02	12	770	770	54862	54862	.0140	.0140

Table 5: Forecasts of Fatalities using EXPSMOOTHING (General Electric) and Unadjusted Data for Calender Year:

January - December

Cyclic Analysis Permitted

Length of Record Months, Years	α	Order	MAD	Lead Time	∑Δ	∑S	∑*S	∑Δ/∑S	∑Δ/∑*S
4,0	.13462	2	142.62	12	-2300	47700	47700	-.0482	-.0482
4,3	.17500	2	133.43	9	36	37720	47700	.0010	.0008
4,6	.21429	2	113.47	6	27	26490	47700	.0010	.0006
4,9	.18750	3	118.06	3	66	13280	47700	.0050	.0014
6,0	.13462	2	166.46	12	-1685	53041	53041	-.0318	-.0318
6,3	.17500	2	143.53	9	-1259	42284	53041	-.0298	-.0234
6,6	.17857	2	128.89	6	4	29284	53041	.0001	.0001
6,9	.18750	3	135.00	3	687	14597	53041	.0471	.0130
8,0	.13462	2	168.22	12	-230	54862	54862	-.0042	-.0042
8,3	.17500	2	158.43	9	-756	43518	54862	-.0174	-.0138
8,6	.17857	2	138.00	6	-136	30075	54862	-.0045	-.0025
8,9	.25000	2	130.11	3	184	15153	54862	.0121	.0034
11,3	.15000	2	149.92	9	-1166	43870	55000	-.0266	-.0212
11,6	.14286	2	140.22	6	-1111	30290	55000	-.0367	-.0202
11,9	.25000	2	150.29	3	353	15010	55000	.0235	.0064

Table 6: Error Analysis of Forecasts of Fatalities using GEXSMO (Univac 1108) and Unadjusted Data for Calender Year:

January - December

Length of Record	Calculated by program: Exponential Smoothing Factors A	B	C	MAD of Forecast Errors	$\sum\Delta$	$\sum S$	$\sum{*}S$	$\sum\Delta/\sum S$	$\sum\Delta/\sum{*}S$
4,0	.07	.55	.15	150.59	-1105	47770	47700	-.0231	-.0231
4,3	.11	.49	.15	155.54	+1365	37720	47700	.0362	.0286
4,6	.10	.55	.12	151.13	+ 407	26490	47700	.0154	.0085
4,9	.09	.55	.14	148.60	+ 225	13280	47700	.0169	.0047
6,0	.12	.37	.15	147.86	-1672	53041	53041	-.0315	-.0315
6,3	.16	.41	.08	143.57	- 215	42284	53041	-.0051	-.0041
6,6	.13	.36	.10	146.52	+ 549	29284	53041	-.0187	.0104
6,9	.13	.34	.10	147.10	+ 679	14597	53041	.0465	.0128
8,0	.25	.33	.01	141.86	- 724	54862	54862	-.0132	-.0132
8,3	.25	.33	.01	149.34	+ 522	43518	54862	.0120	.0095
8,6	.25	.33	.01	146.12	+ 769	30075	54862	.0256	.0140
8,9	.25	.32	.01	144.53	+ 318	15153	54862	.0210	.0058
11,3	.18	.29	.01	152.68	- 969	43870	55000	-.0221	-.0176
11,6	.13	.32	.09	153.32	- 917	30290	55000	-.0303	-.0167
11,9	.13	.31	.09	156.32	- 150	15010	55000	-.0100	-.0027
Three month forecast 1/									
4,0	.07	.55	.15	150.59	- 749	9980	---	-.0751	---
6,0	.12	.37	.15	147.86	- 349	10757	---	-.0324	---
8,0	.25	.33	.01	141.86	- 238	11344	---	-.0210	---

1/ Data picked off from 12 month forecast output

Table 7: Forecasts of Fatalities by Current NHTSA Methods

Year, Month	Forecast Year	Totals by Year Y_{-1}	Y_{-2}	Y_{-3}	4 $\sum_1^3 Y_{-i}$	Totals by Corresponding months m_0	m_{-1}	m_{-2}	m_{-3}	$\sum_1^3 m_{-1}$	m_o/m_{-1}	Method 1 $\hat{Y}_1$	λ (Adjustment Factor)	Method 2 $\hat{Y}_2$	14 $m_o/\sum m_i$	14 * 4	Method 3 $\hat{Y}_3$
4,3	1964 LY	43564	40804	38091	122459	9980	8625	8110	7690	24425	1.1571	50408	90/91 = .98901	49854	.40850	50037	49487
4,6						21210	19004	17962	16872	53858	1.1161	48622	181/182 = .99451	48355	.39396	48244	47979
4,9						34420	31299	29459	27055	87813	1.0997	47907	273/274 = .99635	47732	.39197	48000	47825
6,3	1966	49163	47700	43564	140427	10757	9893	9980	8625	28498	1.0873	53455	90/90 1.00000	53455	.37746	53006	53006
6,6						23757	21733	21210	19004	61947	1.0931	53740	181/181 1.00000	53740	.38350	53854	43854
6,9						38444	34911	34420	31299	100630	1.1012	54138	273/273 1.00000	54138	.38203	53647	53647
8,3	1968 LY	52924	53041	49163	155128	11344	10877	10757	9893	31527	1.0429	55194	90/91 .98901	54587	.35982	55818	55195
8,6						24787	23589	23757	21733	69047	1.0508	55613	181/182 .99451	55308	.35882	55663	55357
8,9						39709	38099	38444	34911	100630	1.0423	55163	273/274 .99635	54962	.35628	55269	55067
11,3	1971	54800	56000	54862	165662	11130	11580	11560	11344	34484	0.9611	52668	90/90 1.00000	52668	.32276	53469	53469
11,6						24710	25130	25520	24787	75437	0.9845	53951	181/181 1.00000	53951	.32756	54264	54264
11,9						39990	40160	40900	39709	120769	0.9958	54570	273/273 1.00000	54570	.33113	54856	54856
12,3	1972 LY	55099	54800	56000	165800	11380	11130	11580	11560	34270	1.0629	58460	90/91 .98901	57818	.34520	57234	56604

Table 8: Error Analysis of Current NHTSA's Methodologies of Fatalities Forecasting and Comparison with EXPSMOOTHING and GEXSMO Methods

												$\sum^*\Delta/\sum^*S$			
													EXPSMOOTHING		
Forecast For	Length of Period	Method 1 Forecast	Method 2 Forecast	Method 3 Forecast	$\sum^*\Delta_1$ Method 1	$\sum^*\Delta_2$ Method 2	$\sum^*\Delta_3$ Method 3	$\sum^*S$	$\sum^*\Delta_1/\sum^*S$	$\sum^*\Delta_2/\sum^*S$	$\sum^*\Delta_3/\sum^*S$	GEXSMO	SAC	SAN	Unadj C
1964	4,3	50408	49854	49487	+2708	+2154	+1787	47700	.0568	.0452	.0375*	.0268	.0121	.0121	.0008+✓
	4,6	48621	48354	47979	+ 921	+ 654	+ 279	47700	.0193	.0137	.0058*	.0085	.0206	.0205	.0006+✓
	4,9	47908	47733	47825	+ 208	+ 33	+ 125	47700	.0044	.0007*✓	.0026	.0047	.0128	.0128	.0014+
1966	6,3	53455	53455	53006	- 414	- 414	- 35	53041	-.0078	-.0078	-.0007*✓	-.0041+	-.0189	-.0200	-.0237
	6,6	53740	53740	53854	- 699	- 699	+ 813	53041	-.0132*	-.0132*	.0153	.0104	.0163✓	+.0172	.0001+✓
	6,9	54138	54138	53647	+1097	+1097	+ 606	53041	.0207	.0207	.0114*	.0128	.0074+	.0123	.0130
1968	8,3	55194	54587	55195	+ 332	- 275	+ 333	54862	+.0061	-.0050*✓	+.0061	.0095	-.0116	-.0087+	-.0138
	8,6	55613	55308	55357	+ 751	+ 446	+ 495	54862	+.0137	+.0081*	+.0090	.0140	.0043	.0094	-.0025+✓
	8,9	55163	54962	55067	+ 301	+ 100	+ 205	54862	+.0055	+.0018*✓	+.0037	.0058	-.0104	.0047	-.0034+
1971	11,3	52668	52668	53469	-2332	-2332	-1531	55000	-.0424	-.0424	-.0278*	-.0176	-.0171	-.0142+✓	-.0202
	11,6	53951	53951	54264	-1049	-1049	- 736	55000	-.0191	-.0191	-.0134*	-.167	-.0095✓	-.0060+✓	-.0201
	11,9	54570	54570	54856	- 430	- 430	- 144	55000	-.0078	-.0078	-.0026*	-.0027	-.0024+	.0059	.0064
Frequency Distribution of Best of 3 NHTSA Methods *									1T	4+1T	7	--	--	--	--
Frequency Distribution of Best of Smoothing Methods +									--	----	--	1	2	3	6
Frequency Distribution of Best of 7 Methods ✓									0	2	1	0	3	2	4
								$\sum$AD	.2168	.1855	.1359	.1336	.1434	.1439	.1070
								MAD	.0181	.0155	.0113	.0111	.0119	.0120	.0089
								Max AD	.0568	.0452	.0375	.0268	.0206	.0206	.0237
								Min AD	.0055	.0007	.0007	.0041	.0024	.0047	.0008

Note: nT = n ties

Table 9: Effect of Lead Time on EXPSMOOTHING Forecasts

Forecast for	Length of Record	$\sum * \Delta / \sum * S$					
		L = 12 EXPSMOOTHING			L = 3		
		SAC	SAN	Unadj C	SAC	SAN	Unadj C
1964	4	-.0456	-.0463	-.0482	-.0254	-.0304	-.0023
1966	6	.0240	-.0332	-.0318	.0475	-.0046	.0009
1968	8	-.0140	.0110	-.0042	-.0332	-.0237	.0029

SAC = Seasonally Adjusted Data with Cyclic Analysis
SAN = Seasonally Adjusted Data with No Cyclic Analysis
Unadj C = Unadjusted Data with Cyclic Analysis

APPENDIX A

Exponential Methods for Analysis and Forecasting of Economic Time Series

From approximately 1960, most methods of time series smoothing analysis have gone toward the use of varying weights based on the age of the data. The simple moving average ignores data beyond a certain age and weights equally each datum that is used. The most popular weighting schemes in vogue today are based on some form of "exponential" smoothing. These schemes use all of the data, with the weight given an individual datum decreasing with increasing age of the datum. The rate of decreasing weight is determined by a parameter called a smoothing constant and most frequently denoted by the Greek letter alpha (α). Depending on the nature of the time series, a variety of models may be used to represent the time series, for example, a constant process model $\xi_t = a_0$, a linear process model $\xi_t = a_0 + a_1 t$, a quadratic process model $\xi(t) = a_0 + a_1 t + a_2 t^2$, etc. In each case there is a process $\xi(t) = \sum_{i=0}^{n} a_i t^i / i!$ which is observed in the presence of noise $x(t) = \xi(t) + \varepsilon(t)$. The number of "degrees of freedom" of these models are respectively 1, 2, 3, etc., and this number is sometimes referred to as the order of the model.

Multiple Smoothing: If a constant model ($\xi_t = a$) is used to represent the time series $x(t) = \xi(t) + \varepsilon(t)$, exponential smoothing for estimating the single coefficient in a constant model is

$$\hat{a}_t = S_t(x) = \alpha x_t + (1-\alpha)\, S_{t-1}(x)$$

For estimating the two coefficients in a linear model, the notion of <u>double</u> smoothing is used. For generality the exponential smoothing for the constant model is called <u>single</u> smoothing and is represented by

$S_t^{[1]}(x)$, i.e., $S_t^{[1]}(x) = S_t(x)$. Then <u>double</u> smoothing is defined as $S_t^{[2]}(x) = S_t^{[1]}(x) + (1-\alpha)\, S_{t-1}^{[2]}(x)$,

Similarly, <u>multiple</u> smoothing of order k is defined by

$S_t^{[k]}(x) = \alpha S_t^{[k-1]}(x) + (1-\alpha)\, S_{t-1}^{[k]}(x)$,

i.e., k^{th} - order smoothing is just simple exponential smoothing applied to the results of (k-1)st - order smoothing of the data.

As an example, in the case of a linear model:

$$\xi_t = a_0 + a_1 t$$

the forecast of the value of the time series τ units after time T is given by

$$\hat{x}_\tau(T) = \hat{a}_0(T) + \tau \hat{a}_1(T)$$

where $\hat{a}_0(T)$ and $\hat{a}_1(T)$ are given by

$$\hat{a}_o(T) = 2\, S_T(x) - S_T^{[2]}(x)$$

$$\hat{a}_1(T) = \frac{\alpha}{1-\alpha}\left[S_T(x) - S_T^{[2]}(x)\right]$$

A convenient, comprehensive basic reference is R.G. Brown, Smoothing, Forecasting, and Prediction of Discrete Time Series, Prentice-Hall, 1963. A summary of the fundamental formulas of multiple smoothing and forecasting are given on pages 142-144, and pages 184ff of this reference.

We remark at this time that in using multiple smoothing methods two problems arise: (a) what order of smoothing should be used? (b) what value of the smoothing constant should be employed? A third, but minor, problem is the determination of starting values.

The mathematical basis for the EXPSMOOTHING computer program is founded on the concepts of multiple smoothing, and utilizes the formulas presented in Brown's book.

A similar but different approach forms the basis of the UNIVAC GEXSMO program. This generalized exponential smoothing program is based on the work of P.R. Winters, "Forecasting Sales by Exponentially Weighted Moving Averages," Management Science, April, 1960. The essentials of this computer program are given in Appendix B under the section UNIVAC 1108.

APPENDIX B

Available Computer Programs at NBS for Time Series Analysis

Several time-sharing interactive systems were immediately available at NBS. These were DIALCOM, Computer Sciences Corporation Conversational Executive and General Electric Mark II. In addition the Large Scale Systems STAT-PACK for the UNIVAC 1108 tests several routines for time series analyses. A summary of the pertinent program libraries of each of these systems follows.

DIALCOM: Moving Average (A Simple Moving average)
Autocovariance
Cross Covariance
Smoothed Series (A weighted moving average)
Seasonal Index and Cyclical Movement

Note: No expontential smoothing method is available and none of the programs have an extrapolation feature.

Computer Sciences Conversational Executive (CSCX Basic Library): This has a Triple Exponential Smoothing routine called ***SMOOTH. In addition to smoothing, and listing the smoothed values and differences, the routine can produce a plot of the observed values and the smoothed values (on the same graph) and an extrapolation feature is also available.

There is no routine available for first or second order smoothing and there is no automatic feature such as a mean absolute deviation. The value of the smoothing constant must be supplied by the user.

General Electric (Mark II): The EXPSMOOTHING routine is a very comprehensive routine which performs a sequence of exponential smoothing operations for a number of alpha values for each order of smoothing: 1, 2, 3. The routine then selects that combination of alpha and order of smoothing that has minimum mean absolute deviation per data point. A forecast (extrapolation) option is also available, along with a plot. A variety of output options is possible.

The routine also permits the users to override several automatic features of the routine. For example, the user has the capability to specify cycle length of known periodic data and also to specify the values of the smoothing constants, but not the order of smoothing.

A summary of the essential details of this particular routine is given in Appendix C.

UNIVAC 1108: The Large Scale Systems STAT-PACK for the UNIVAC 1108 lists several routines for time series analysis. These are UNIVAC system subroutines and are not necessarily National Bureau of Standards routines, but they are available for use on the NBS UNIVAC 1108. The most appropriate subroutine for the problem at hand appears to be GEXSMO - Generalized

Exponential Smoothing.* The routine

1. Produces the seasonally adjusted series point $\tilde{S}_t$
2. Computes a seasonality factor F_t to be used in the next cycle
3. Computes the next trend value R_t
4. Produces a forecast for the next period,** $S_{t,1}$
5. Compares the forecast with the actual to get the forecast error

$$e_{t,1} = S_{t+1} - S_{t,1}$$

where S_{t+1} is the (t+1)st time series data point.

When all the data are used, L forecasts for <u>future periods</u> may be produced by using the equation

$$S_{N,K} = (\tilde{S}_N + KR_N)\ F_{N-L+K} \quad K = 1, 2, \ldots, L \text{ (Linear increasing)}$$

where L is the periodicity of the data, e.g., if the data is by month, L = 12 and N is the number of elements in the time series.

Three exponential smoothing factors, A, B, and C all between 0 and 1 are used respectively to produce the seasonally adjusted series, to adjust the seasonality factors and to adjust the trend. These smoothing factors and initial values may be input or may be calculated by the routine.

---

*The notation used here is that employed in the UNIVAC Programmers Reference Manual. The mathematical background for the method as given in Winters, P.R.: Forecasting Sales by Exponentially Weighted Moving Averages, <u>Management Science</u>, April, 1960.

**In this context "next period" refers to the next time point.

The output of the GEXSMO routine is listed under the headings:

ORIGINAL SERIES	SEASONALLY ADJ. SERIES	FORECASTED SERIES	SEASONALITY	TREND	ERROR

and this is followed by

MEAN ABSOLUTE DEVIATION OF FORECAST ERRORS _______________

PERCENTAGE ERROR __________________

We mention the availability of other UNIVAC time series routines:

1. MOVAVG - Moving Averages. This subroutine computes the smoothing coefficients in a polynomial model for a time series and smooths the given time series.
2. SEASHI - Shiskin's Seasonality Factors. This subroutine produces seasonality factors from a time series, and smooths, detrends and deseasonalizes the series.
3. WEMAV - Weighted Moving Averages. This subroutine eliminates a trend from a time series by a weighted moving average.
4. TRELS - Trend Analysis by Least Squares. This subroutine removes trends in a time series with a general linear model, using least squares.

5. VADIME - Variate Difference Method. This subroutine estimates the variance of the random component in a time series and determines a lower limit for the degree of the polynomial which can be used in approximating the trend.

6. TSFARG - Autoregressive Model. This subroutine obtains the least square coefficients in an autoregressive time series model and produces forward forecasts from the model.

The documentation of the detail of each of these seven programs is very adequately done in the UNIVAC Programmer's Reference Manual for STAT-PACK and is not repeated here.

APPENDIX C

Details of EXPSMOOTHING Routine

The EXPSMOOTHING routine is a very comprehensive routine which performs a sequence of exponential smoothing operations for a number of alpha values for each order of smoothing: 1, 2, 3. The routine then selects that combination of alpha and order of smoothing that has minimum mean absolute deviation per data point. A forecast (extrapolation) option is also available, along with a plot. A variety of output options is possible.

The routine also permits the users to override several automatic features of the routine. For example, the user has the capability to specify cycle length of known periodic data and also to specify the values of the smoothing constants, but not the order of smoothing.

For the convenience of the reader we present a summary of the essentials of the General Electric Time Sharing EXPSMOOTHING Routine. The discussion is based on experience in using the routine and explanations of this routine from two User's Guides* prepared by the

---

*Mark II User's Guide: Statistical Analysis System, General Electric Company, December, 1971 (5707.01).

GETSA$: Mark I Marketing and Economic Forecasting, A Hands-On Users Guide, General Electric Company, 1969. (906329) the General Electric Company, and is prepared in the sequence provided by the version in use in December, 1972. For additional details on program operation the reader is referred to the user's guides. The time series data to be subject to the EXPSMOOTHING routine is called RAW DATA by the program. The RAW DATA may be adjusted initially by what is called a BASE SERIES. This series permits one to remove known distortions that one would not expect to occur at future times. The base series may be used to take into account discontinuities, human judgment or the results of statistical analyses of the data. The base series values are subtracted from the raw data before an analysis is made and later added back into the forecast.

To operate the program, in addition to entering the raw data and the base series, one has to indicate two times, the LEAD TIME (L) and the FORECAST HORIZON (H) as well as a choice of six output options.

Lead time is the term used to describe the time (in periods) between the last piece of data one has as data as input and the period one is most interested in estimating correctly. The forecast will be optimized according to the lead time that is entered.

The forecast horizon indicates how many periods beyond the last data point will be forecasted, but has nothing to do with the optimization. For example, if the raw data is reported monthly and the user wishes to forecast six months in the future, accurately, and also wishes to take a casual glance at the forecast one year ahead, a lead time of six and a horizon time of 12 would be indicated.

The output options are cumulative, so that option code 5 gives all options. The codes for the output options are:

0 = Simple forecast table

1 = Raw data residue

2 = Raw data list

3 = Base series list and cyclic analysis

4 = Cyclic forecast

5 = Trend and error analysis and composite forecast.

A graphical display of the raw data and forecasted (smoothed) and extrapolated data may also be obtained.

The following discussion is based on maximum output (Code 5). The output indicates whether the smoothing constant was supplied or provided automatically by the programs and

a. NUMBER OF RAW DATA POINTS

b. NUMBER OF BASE DATA POINTS

c. FORECAST HORIZON

d. LEAD TIME.

Next, the RAW DATA are printed out.

Then comes DETERMINATION OF THE OPTIMUM NUMBER OF DATA POINTS PER CYCLE. The entries are in columns K and ERR(K), where K refers to cycle length as expressed in data periods, and where K runs from 1 to half the number of data points. The ERR(K) shows a measure* of error associated with each cycle length. The GETSA$ user's guide states:

> The lower the error, the stronger is the tendency for the data to conform to a cyclic pattern. The absolute value of the error term has little significance for interpretation, but the relative value shows which cycle length fits best and that length (K value) is chosen as the optimum length and subsequent calculations assume a cycle of that length.... The optimum number of periods per cycle will be chosen based on minimum error. That cycle length is assumed for the remainder of the analyses.

If min ERR(K) occurs for K = 1, the routine will indicate that the data is not cyclic.

Then the output indicates THE OPTIMUM PERIOD FOR THE CYCLE and prints out the corrections that are used to remove the cyclic effects, under the headings T and C(T)**.

---

*We have been unable to determine the exact nature of this measure.

**We have been unable to determine the exact nature of this correction.

Next there appears TABULATION OF THE ACTUAL (UNSMOOTHED) RESIDUE. The residue at this stage contains the corrections for base series and cyclic effects. At this point the

Residue (T) =

$$\text{Raw Data Residue (T)} - \sum_{\tau-1}^{T} C(\tau-1),$$

where C(0) = 0.

The output then indicates the results of EXPONENTIAL SMOOTHING APPLIED TO RESIDUE: under the headings

ALPHA	TYPE OF SMOOTHING	MEAN ABSOLUTE DEVIATION PER DATA POINT

If the alpha values are not supplied but are determined by the program, the program will select* eight (8) values of alpha and do exponential smoothings of orders 1, 2, and 3 for each alpha. The minimum absolute deviation per data point for this set of 24 smoothings then determines the alpha value and order of smoothing that will be employed in the so-called FORECAST section of the program. The output headings at this stage are:

OPTIMUM ALPHA	TYPE OF SMOOTHING	MEAN ABSOLUTE DEVIATION PER DATA POINT

followed by ------------------------FORECAST-------------

TIME	RESIDUE	COMPOSITE	ACTUAL	ERROR

---

*If α is not forced, program will select 8 values of α based on $\alpha = \frac{1}{8} \; \frac{2}{L+1}$. The leading term is always $\alpha_1 = .01$. The second term is $\alpha_2 = \min(.05, \alpha)$.

If $\alpha_2 = .05$, $\alpha_3 = \alpha$, $\alpha_4 = 2\alpha$, ..., $\alpha_8 = 6\alpha$.

If $\alpha_2 = \alpha$, $\alpha_3 = 2\alpha$, $\alpha_4 = 3_\alpha$, ..., $\alpha_8 = 7\alpha$.

Under TIME the initial entry is at time 5 + L, where L is the lead time. Several observations are required before a smoothed average can be calculated and this routine uses five.

The RESIDUE indicated is the smoothed residue.

Under COMPOSITE appear the results of putting back in the reverse corrections for cyclic effects and base series.

Under ACTUAL are listed the values that were originally listed as RAW DATA.

Under ERROR appear the results of ACTUAL minus COMPOSITE. The entries under RESIDUE and COMPOSITE beyond the last ACTUAL data point are based on the exponential smoothing forecast formulas* of orders 1, 2, 3:

$$F1_{t+L} = S1_t = CED1_t$$

$$F2_{t+L} = CED2_t + L(C2_t)$$

$$F3_{t+L} = CED3_t + L(C3_t) + 1/2\ L^2(RC3_t)$$

The values that are actually used in the extrapolation are listed under the headings:

S1	S2	S3
CED1	CED2	CED3
C2	C3	RC3

The values shown are for the last raw data point.

---

*These and other formulas used in the exponential smoothing process can readily be transformed into the notation used in Brown's book in pages 142-144 and pages 184ff.

This is followed by HORIZON FORECAST BEGINS AT TIME ____ and LEAST SQUARES FIT and MEAN and VARIANCE.

The LEAST SQUARES FIT is simply an equation describing the straight line trend which best approximates (in the sense of least squares) movement of the raw data for all the data.

At this time the computer will ask WANT A PLOT? and WANT DATA STORED? If a plot is desired, this routine will give on one graph plots of the values listed under FORECAST COMPOSITE and ACTUAL.

FORM NBS-114A (1-71)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET	1. PUBLICATION OR REPORT NO. NBSIR - 73-138	2. Gov't Accession No.	3. Recipient's Accession No.

4. TITLE AND SUBTITLE Time Series Forecasting of Highway Accident Fatalities	5. Publication Date 6. Performing Organization Code
7. AUTHOR(S) Dr. Alexander R. Craw	8. Performing Organization NBSIR - 73-138
9. PERFORMING ORGANIZATION NAME AND ADDRESS NATIONAL BUREAU OF STANDARDS DEPARTMENT OF COMMERCE WASHINGTON, D.C. 20234	10. Project/Task/Work Unit No. 4314496 11. Contract/Grant No.
12. Sponsoring Organization Name and Address National Highway Traffic Safety Administration Department of Transportation Washington, D.C. 20590	13. Type of Report & Period Covered 14. Sponsoring Agency Code

15. SUPPLEMENTARY NOTES

16. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)

Using twelve years of time series data on highway fatalities, the methodology currently employed by the National Highway Traffic Safety Administration (NHTSA) to forecast the annual (calendar year) total of highway accident fatalities were compared with those obtained by several computer routines based on exponential smoothing techniques and available at the National Bureau of Standards. The use of unadjusted and seasonally adjusted data was also examined.

It is found that there is no coercive evidence to lead to abandoning the present NHTSA methods in favor of readily available computer routines based on exponential smoothing methods.

Of the methods examined in this study, the best results were obtained with the EXPSMOOTHING routine using unadjusted fatality data.

17. KEY WORDS (Alphabetical order, separated by semicolons)
Exponential smoothing; fatalities; highway accidents; time series

18. AVAILABILITY STATEMENT	19. SECURITY CLASS (THIS REPORT)	21. NO. OF PAGES
[X] UNLIMITED.	UNCLASSIFIED	
[ ] FOR OFFICIAL DISTRIBUTION. DO NOT RELEASE TO NTIS.	20. SECURITY CLASS (THIS PAGE) UNCLASSIFIED	22. Price