A MATHEMATICAL THEORY OF SEASONALS

By

STATISTICAL DEPARTMENT, DETROIT EDISON CO.

The graph of any time series may be assumed to be a compound curve which is dependent upon the following factors:

Secular trend,	$f(x)$
Cycle,	$c(x)$
Seasonal	$s(x)$, and
Residual errors,	ϵ_x

If we designate the xth term of the observed time series by ${}_oy_x$, we have that

$$(1) \qquad {}_oy_x = f(x)\cdot c(x)\cdot s(x) + \epsilon_x$$

It also follows that the standard error, based on our hypothesis, is

$$(2) \qquad \sigma_\epsilon = \sqrt{\frac{\sum \epsilon_x^2}{N}}$$

In making predictions, we desire that the standard error of estimate be a minimum, and this requires that $\sum \epsilon^2$ be also a minimum.

In dealing with data covering a period of years, i. e. $12n$ months, we observe that

$$\sum \epsilon^2 = \left[{}_0y_1 - f(1)\cdot c(1)\cdot s(1)\right]^2$$

$$+ \left[{}_0y_2 - f(2)\cdot c(2)\cdot s(2)\right]^2$$

$$\text{---------------------}$$

$$+ \left[{}_0y_{12} - f(12)\cdot c(12)\cdot s(12)\right]^2$$

$$+ \left[{}_0y_{13} - f(13)\cdot c(13)\cdot s(1)\right]^2$$

$$+ \left[{}_0y_{14} - f(14)\cdot c(14)\cdot s(2)\right]^2$$

$$\text{---------------------}$$

$$+ \left[{}_0y_{12n-11} - f(12n-11)\cdot c(12n-11)\cdot s(1)\right]$$

$$+ \left[{}_0y_{12n-10} - f(12n-10)\cdot c(12n-10)\cdot s(2)\right]$$

$$\text{------------------------------------}$$

$$+ \left[{}_0y_{12n} - f(12n)\cdot c(12n)\cdot s(12)\right]$$

Let us now find the values of $s(1)$, $s(2)$ . . . $s(12)$ that will minimize the standard error of estimate. Placing the partial derivative of $\sum \epsilon^2$ with respect to $s(1)$ equal to zero, yields

$$\frac{\partial}{\partial s1} = 2\left[{}_0y_1 - f(1)\cdot c(1)\cdot s(1)\right]\left[-f(1)c(1)\right]$$

$$+ 2\left[{}_0y_{13} - f(13)\cdot c(13)\cdot s(1)\right]\left[-f(13)\cdot c(13)\right]$$

$$\text{------------------------------------}$$

$$+ 2\left[{}_0y_{12n-11} - f(12n-11)\cdot c(12n-11)\cdot s(1)\right]\left[-f(12n-11)\cdot c(12n-11)\right] = 0$$

Solving

$$s(1) = \frac{\sum^{(1)} {}_0y_x\, f(x)\cdot c(x)}{\sum^{(1)} f^2(x)\cdot c^2(x)}$$

where we understand that $\sum^{(1)} {}_0y_x \cdot f(x) \cdot c(x)$ means the sum of the products of ${}_0y_x$, $f(x)$ and $c(x)$ taken from the *first* month of each year, and similarly for $\sum^{(1)} f^2(x) \cdot c^2(x)$

The partial derivative with respect to $s(2)$ yiel

$$s(2) = \frac{\sum^{(2)} {}_0y_x \cdot f(x) \cdot c(x)}{\sum^{(2)} f^2(x) \cdot c^2(x)}$$

and in fact

$$s(i) = \frac{\sum^{(i)} {}_0y_x \cdot f(x) \cdot c(x)}{\sum^{(i)} f^2(x) \cdot c^2(x)} \tag{3}$$

Thus the seasonal for July is a function only of the various July values of the observed series, the secular trend and the cycle factors.

Since both $f(x)$ and $c(x)$ are smooth functions, it follows that their product, which we shall designate by $\psi(x)$, represents a smooth function which is merely that part of the time series which would remain if the accidental and seasonal fluctuations were eliminated. The formula for the seasonal index for the i th month may therefore be written

$$s(i) = \frac{\sum^{(i)} {}_0y_x \cdot \psi(x)}{\sum^{(i)} \psi^2(x)} \tag{4}$$

At this point we may recall the fact that in fitting a curve of the type $y = k\psi(x)$ to observed data by the *Method of Least Squares*,

$$k = \frac{\sum {}_0y_x \cdot \psi(x)}{\sum \psi^2(x)}$$

whereas if the *Method of Moments* be employed

$$k = \frac{\sum {}_0y_x}{\sum \psi(x)}$$

Experience in various statistical applications demonstrates that the two methods yield approximately the same results. Borrowing from this experience, we shall choose the simpler form and write instead of formula (4)

$$(5) \qquad s(i) = \frac{\sum\limits_0^{(i)} y_x}{\sum\limits^{(i)} \psi(x)}$$

So far as theoretical considerations are concerned (4) may be superior to (5), but the fact that the latter formula enables us to obtain seasonals by a method far simpler than would result by using formula (4), requires that we choose (5) in preference to (4). Ordinarily the difference in results obtained by using both formulae is less than one-half of one per cent.

Verbally, formula (5) states merely that *the seasonal index for any month is the ratio of the total of the variates for the month in question to the total that would have been experienced if neither accidental nor seasonal influences were present.*

We now are forced to find a simple method of obtaining values of $\sum\limits^{(i)} \psi(x)$.

Let T_{i-3}, T_{i-2}, T_{i-1}, T_i, T_{i+1}, T_{i+2}, and T_{i+3} denote the total production for seven consecutive years. If we assume that the effect of both seasonal influences and accidental or residual fluctuations is to shift the production from one month to another, but nevertheless *to leave the total production for each year practically unchanged,* then a smooth curve passing over the seven year period, and preserving the annual totals, may be assumed to afford a representation of $\psi(x)$. We, therefore, determine the equation of a parabola of the sixth degree in such a manner that the areas under this curve for seven equidistant unit intervals are equal respectively to $T_{i-3}, T_{i-2}, \cdots T_0, \cdots T_{i+2}, T_{i+3}$ Fitting six degree parabolae to successive seven year intervals it is possible to deal with a time series of any length.

By adding together the interpolated values for all the January values of $\psi(x)$, and similarly for the other months, we can show that

$$(6) \quad \sum^{(i)} \psi(x) = c_{1:i}\, T_1 + c_{2:i}\, T_2 + c_{3:i}\, T_3 + c_{4:i}\left[T_4 + T_5 + \cdots T_{n-3}\right]$$
$$+ c_{5:i}\, T_{n-2} + c_{6:i}\, T_{n-1} + c_{7:i}\, T_n$$

where the values of the coefficients are as given in Table I.

In order to compare the efficiency of this method with another method of computing seasonals, it is necessary that each formula be tried out on some series for which the true values of the seasonal indices are known. We know in advance, of course, that there exist many satisfactory methods of obtaining seasonals, but we also desire to know something about the amount of time that each method requires as well as their relative accuracy.

The theoretical series, on which we shall try out two methods of computing seasonals, is built up from data taken from an article, "Statistical Analysis and Projection of Time Series," written and published by the statistical division of the American Telephone and Telegraph Company. After eliminating from the *Production of Pig Iron* series both trend and seasonal influences, the factors of Table II remained. We shall consider these, therefore, as the combination of "Cycle and residual" factors.

Although smoothing this data by a proper mathematical formula would eliminate the residual errors, nevertheless such procedure would introduce a bias in favor of the formula for computing seasonals that is proposed in this paper. The reason for this bias lies in the fact that most smoothing formulae are developed on the assumption that the smoothed ordinate lies on a parabola of a chosen degree, and since a similar assumption was made in our theory, it is evident that the proposed method will benefit most by employing a parabolic smoothing formula in obtaining the hypothetical cycle series.

For this reason the data of Table II, with additional data for one year on either side, was given to a draftsman with instructions to

(1) Plot the data of Table II

(2) draw free hand a *smooth* curve that to his mind best represented the general run of the data

(3) read off from his curve the approximate value of the smoothed statistics.

The data of Table III resulted.

In essential agreement with the American Telephone and Telegraph article, we shall assume a linear trend, the value for the first month being 1511 and the monthly increment 8. The product of trend by cycle produces the theoretical values of $\psi(x)$ presented in Table IV.

TABLE I

Constants for computing seasonal indices

i	$c_{1:i}$	$c_{2:i}$	$c_{3:i}$	$c_{4:i}$	$c_{5:i}$	$c_{6:i}$	$c_{7:i}$
1	.12530	.07897	.08392	.083333	.08259	.08959	.03963
2	.11822	.07914	.08389	.083333	.08269	.08849	.04757
3	.11094	.07955	.08382	.083333	.08283	.08723	.05563
4	.10345	.08018	.08373	.083333	.08299	.08590	.06375
5	.09577	.08104	.08361	.083333	.08315	.08456	.07187
6	.08792	.08208	.08347	.083333	.08331	.08327	.07995
7	.07995	.08327	.08331	.083333	.08347	.08208	.08792
8	.07187	.08456	.08315	.083333	.08361	.08104	.09577
9	.06375	.08590	.08299	.083333	.08373	.08018	.10345
10	.05563	.08723	.08283	.083333	.08382	.07955	.11094
11	.04757	.08849	.08269	.083333	.08389	.07914	.11822
12	.03963	.08959	.08259	.083333	.08392	.07897	.12530

TABLE II

Cycle and Residual Series for Pig Iron Production

	1904	1905	1906	1907	1908	1909
January	–37.5	12.4	23.0	24.2	–43.3	– 7.9
February	–13.8	5.8	18.2	20.2	–36.4	– 7.8
March	–10.4	14.2	20.2	17.6	–40.8	–13.7
April	– .7	15.9	18.1	19.7	–42.0	–15.6
May	– 4.2	14.6	17.5	21.9	–43.0	–10.5
June	–15.2	10.4	15.0	23.1	–42.0	– 3.3
July	–27.4	7.0	16.8	23.9	–35.5	4.9
August	–25.3	11.1	9.8	21.6	–30.6	10.1
September	–12.5	15.4	12.7	19.0	–25.8	18.0
October	–12.1	18.6	20.2	21.4	–24.1	22.6
November	– 5.6	22.1	23.8	– 1.6	–19.0	24.2
December	1.1	20.7	24.9	–34.5	–12.1	27.0
	1910	**1911**	**1912**	**1913**	**1914**	**1915**
January	26.7	–17.7	– 8.0	19.5	–22.0	–35.9
February	21.5	–11.0	– 1.0	15.7	–16.7	–27.8
March	19.5	– 5.5	– .1	10.6	–10.5	–25.0
April	15.8	– 8.2	1.4	13.0	–10.8	–20.1
May	9.4	–17.9	5.4	13.9	–19.7	–16.3
June	8.2	–17.9	7.0	10.6	–21.9	– 6.9
July	2.5	–17.8	5.5	7.7	–20.3	.0
August	– 1.3	–13.6	8.1	5.1	–20.6	6.5
September	– 2.5	–10.1	7.1	4.7	–23.8	10.5
October	– 6.5	–10.3	11.0	.7	–33.6	15.1
November	–10.7	–10.5	12.8	– 7.8	–39.3	16.1
December	–18.1	– 9.8	17.8	–19.3	–40.6	21.0

TABLE III

Per Cent Cycle Series for Theoretical Distribution

	1904	1905	1906	1907	1908	1909
January	–38.2	6.6	15.5	16.3	–37.5	–12.2
February	–36.7	7.5	15.8	16.2	–38.1	– 7.1
March	–32.4	9.0	16.0	16.0	–39.1	– 2.4
April	–22.4	10.0	16.2	14.6	–39.4	2.1
May	–18.7	11.0	16.4	14.1	–39.4	5.0
June	–13.5	12.0	16.6	12.7	–38.5	10.0
July	– 9.8	13.1	16.6	11.6	–38.0	12.7
August	– 6.4	13.6	16.6	8.0	–36.7	15.6
September	– 3.7	14.1	16.7	5.0	–33.4	17.2
October	.0	14.5	16.6	.0	–29.1	18.4
November	2.3	14.8	16.5	–17.4	–23.8	19.6
December	4.4	15.0	16.5	–35.0	–17.0	20.0
	1910	**1911**	**1912**	**1913**	**1914**	**1915**
January	20.6	–12.3	– 7.2	10.8	–14.0	–32.3
February	20.6	–13.4	– 5.0	10.7	–17.0	–28.1
March	20.3	–13.6	– 3.0	10.6	–20.0	–23.6
April	19.5	–13.7	.0	10.0	–23.0	–18.1
May	17.9	–13.7	2.7	9.3	–24.7	–13.0
June	16.4	–13.6	4.8	8.5	–27.5	– 7.4
July	12.4	–13.4	6.9	7.1	–29.8	– 2.8
August	7.1	–12.8	7.9	6.0	–32.0	3.5
September	.0	–12.0	8.5	3.6	–33.5	9.0
October	– 6.6	–11.2	9.5	.0	–35.0	13.8
November	– 8.0	– 9.6	10.0	– 5.0	–36.0	15.7
December	–10.9	– 7.9	10.6	–10.0	–35.0	17.5

TABLE IV

Theoretical Trend and Cycle Series, $\psi(x)$

1904	1905	1906	1907	1908	1909
934	1713	1967	2092	1184	1748
962	1736	1981	2100	1178	1857
1032	1769	1994	2105	1164	1959
1191	1794	2007	2089	1163	2057
1254	1819	2020	2089	1168	2124
1342	1845	2032	2073	1190	2234
1406	1872	2042	2061	1205	2298
1467	1889	2051	2003	1235	2366
1517	1907	2062	1956	1305	2408
1583	1922	2070	1871	1395	2443
1628	1937	2077	1552	1505	2477
1669	1949	2087	1227	1646	2495

1910	1911	1912	1913	1914	1915
2517	1914	2115	2632	2125	1738
2527	1897	2173	2638	2058	1851
2530	1900	2226	2644	1990	1973
2523	1905	2303	2639	1921	2122
2498	1912	2373	2631	1885	2261
2476	1921	2430	2620	1820	2414
2400	1932	2488	2595	1768	2542
2295	1952	2519	2577	1718	2715
2151	1977	2542	2527	1686	2868
2017	2002	2574	2447	1653	3003
1994	2046	2595	2332	1633	3063
1938	2092	2618	2217	1663	3120

TABLE V

Theoretical Seasonal Factors

January	.99	May	1.04	September	.98
February	.93	June	.98	October	1.04
March	1.05	July	.98	November	.99
April	1.02	August	1.00	December	1.00

By multiplying the data of Table IV by the seasonals of Table V, a theoretical series would be obtained which would comprise the elements of trend, cycle and seasonal—lacking only chance or residual errors.

In order to obtain a series of chance factors that might serve as residual error factors, sixty cards were marked with integers totaling 1200. The cards were distributed, after shuffling, into twelve piles of five cards each, and the totals of each pile noted. These were taken as the residual factors for the first year, and the process was repeated for the following years. The chance factors of Table VI resulted.

Making allowance for residual errors as well as the seasonal factors, we obtain finally the theoretical series which we shall attempt to analyze, Table VII.

If the various methods of analyzing time series are sound, they should be able to break up this series into its elementary components—trend, cycle, seasonal and residual errors. A comparison of the results by different methods should indicate to some extent their respective merits. In attacking the ordinary observed time series by different methods and comparing results the difficulty is to tell, when all has been done, which of the methods is best. Unfortunately, if they disagree, we do not know which one is nearest the truth. Our theoretical series, however, enables us to compare results obtained by different methods, since we know the answers in advance, and also will serve students as a detailed example of time series synthesis.

TABLE VI

Residual Factors

	1904	1905	1906	1907	1908	1909
January	98	98	98	98	106	99
February	91	98	101	95	96	106
March	105	103	98	94	105	89
April	100	108	99	98	102	99
May	103	100	101	94	99	97
June	94	94	84	103	100	94
July	91	114	102	102	105	103
August	116	93	90	108	108	102
September	98	102	104	118	100	106
October	95	95	107	100	94	104
November	99	84	112	96	90	97
December	110	111	104	94	95	104
	1910	1911	1912	1913	1914	1915
January	96	102	98	98	97	98
February	107	99	102	95	93	89
March	91	95	97	92	105	92
April	110	96	106	104	104	83
May	104	109	98	104	110	108
June	102	92	108	105	97	113
July	94	92	109	91	91	103
August	98	103	98	98	96	98
September	100	102	104	96	100	105
October	105	103	94	112	102	107
November	95	107	93	108	98	102
December	98	100	93	97	107	102

TABLE VII

Theoretical Series

	1904	1905	1906	1907	1908	1909
January	906	1662	1908	2030	1242	1714
February	814	1582	1860	1855	1052	1831
March	1138	1913	2052	2077	1283	1831
April	1215	1976	2027	2088	1210	2077
May	1343	1892	2122	2043	1203	2143
June	1236	1700	1672	2093	1166	2058
July	1254	2092	2041	2060	1240	2320
August	1702	1757	1846	2163	1334	2413
September	1457	1906	2102	2262	1279	2502
October	1564	1899	2304	1946	1364	2643
November	1596	1611	2303	1475	1341	2378
December	1836	2163	2170	1153	1564	2595
Total	16061	22153	24407	23245	15278	26505
	1910	1911	1912	1913	1914	1915
January	2392	1933	2052	2554	2041	1687
February	2514	1746	2061	2330	1780	1532
March	2417	1895	2267	2554	2194	1906
April	2830	1865	2490	2800	2037	1796
May	2702	2167	2419	2845	2156	2539
June	2475	1732	2571	2696	1730	2674
July	2211	1742	2657	2314	1577	2566
August	2249	2011	2469	2525	1649	2661
September	2108	1976	2591	2377	1652	2952
October	2203	2144	2516	2850	1753	3342
November	1875	2168	2389	2494	1585	3093
December	1899	2092	2435	2150	1779	3182
Total	27875	23471	28917	30489	21933	29930

To obtain the values of the seasonal factors by means of formula (5) and Table I we need only observe that for the theoretical series

$$
\begin{aligned}
T_1 &= 16061 \\
T_2 &= 22153 \\
T_3 &= 24407 \\
T_4 + T_5 + \dots T_9 &= 145291 \\
T_{10} &= 30489 \\
T_{11} &= 21933 \\
T_{12} &= 29930
\end{aligned}
$$

Consequently we have

TABLE VIII

Seasonals by Interpolation Method

Month	$\Sigma_0 y_x$	$\Sigma \psi(x)$	s
January	22121	23587	.938
February	20957	23693	.885–
March	23527	23801	.988
April	24411	23911	1.021
May	25574	24023	1.065–
June	23803	24135	.986
July	24074	24246	.993
August	24779	24358	1.017
September	25164	24468	1.028
October	26528	24576	1.079
November	24308	24682	.985
December	25018	24785	1.009
Total	290264	290265	11.994

It is interesting to compare the seasonals of Table VIII with the corresponding set obtained by the method of "link relatives." The following table presents the series of link relatives for the theoretical series of Table VII.

TABLE IX

Link Relatives for the Series of Table VII

	1904	1905	1906	1907	1908	1909
January	.898	.952	.975	.914	.847	1.068
February	1.398	1.209	1.103	1.120	1.220	1.000
March	1.068	1.033	.988	1.005	.943	1.134
April	1.105	.957	1.047	.978	.994	1.032
May	.920	.899	.788	1.024	.969	.960
June	1.015	1.231	1.221	.984	1.063	1.127
July	1.357	.840	.904	1.050	1.076	1.040
August	.856	1.085	1.139	1.046	.959	1.037
September	1.073	.996	1.096	.860	1.066	1.056
October	1.020	.848	1.000	.758	.983	.900
November	1.150	1.343	.942	.782	1.166	1.091
December	.905	.882	.935	1.077	1.096	.922
	1910	**1911**	**1912**	**1913**	**1914**	**1915**
January	1.051	.903	1.004	.912	.872	.908
February	.961	1.085	1.100	1.096	1.233	1.244
March	1.171	.984	1.098	1.096	.928	.942
April	.955	1.162	.971	1.016	1.058	1.414
May	.916	.799	1.063	.948	.802	1.053
June	.893	1.006	1.033	.858	.912	.960
July	1.017	1.154	.929	1.091	1.046	1.037
August	.937	.983	1.049	.941	1.002	1.109
September	1.045	1.085	.971	1.199	1.061	1.132
October	.851	1.011	.950	.875	.904	.925
November	1.013	.965	1.019	.862	1.122	1.029
December	1.018	.981	1.049	.949	.948	

From the above we obtain the following:

TABLE X

Link Relative Seasonal Indices

Months	(1) Medians	(2) Chain Relatives	(3) (2) Adjusted	(4) Seasonal Indices
January	.913	100.0	100.0	97.5
February	1.112	91.3	91.3	89.0
March	1.019	101.5	101.4	98.8
April	1.024	103.5	103.3	100.7
May	.934	105.9	105.7	103.0
June	1.010	98.9	98.7	96.2
July	1.043	99.9	99.7	97.2
August	1.020	104.2	103.9	101.3
September	1.064	106.3	106.0	103.3
October	.914	113.1	112.7	109.9
November	1.024	103.4	103.0	100.4
December	.949	105.9	105.4	102.7
January		100.5	100.0	

The following exhibit of the results obtained by the two methods is interesting.

TABLE XI

Comparison of Interpolation and Link Relative Methods

Months	Actual Values	Interpolation Method		Link Relative Method	
		Seasonal	Error	Seasonal	Error
January	.990	.938	–.052	.975	–.015
February	.930	.885	–.045	.890	–.040
March	1.050	.988	–.062	.988	–.062
April	1.020	1.021	.001	1.007	–.013
May	1.040	1.065	.025	1.030	–.010
June	.980	.986	.006	.962	–.018
July	.980	.993	.013	.972	–.008
August	1.000	1.017	.017	1.013	.013
September	.980	1.028	.048	1.033	.053
October	1.040	1.079	.039	1.099	.059
November	.990	.985	–.005	1.004	.014
December	1.000	1.009	.009	1.027	.027

The mean deviations and the standard deviations of the two methods show that both methods are about equally effective. This advantage of the interpolation method is scarcely worth mentioning. Nevertheless, the fact that the results are obtained with but a trivial amount of labor is important.

	Mean Deviation of Errors	Standard Deviation of Errors
Interpolation Method	.0269	.0337
Link Relative Method	.0277	.0338