Serdica
Bulgariacae mathematicae publicationes

Сердика
Българско математическо списание

SERDICA — Bulgaricae
mathematicae publicationes
18 (1992) 215–223

BACKWARD ERROR ANALYSIS OF LU-DECOMPOSITION FOR PENTADIAGONAL MATRICES

P. Y. YALAMOV

ABSTRACT. Systems with pentadiagonal matrices are often met in practice when solving differential equations numerically. This paper uses the method proposed in [3] to study the round-off error propagation of LU-decomposition for linear systems with pentadiagonal matrices. The results is that the equivalent perturbations of the inputs are relatively small for well-conditioned problems. Backward analysis needs much less computational time than forward analysis if we want to estimate the round-off errors numerically.

1. Introduction. Linear systems with pentadiagonal matrices arise often when solving differential equations numerically. In this paper we use the method proposed in [3] to study the LU-decomposition for linear systems with pentadiagonal matrices with respect to round-off errors. The method is based on the use of the dependence graph of the algorithm and its parallel forms (see [2]). The notion of equivalent pertubation is introduced for every piece of data (input, intermediate and output) in contrast to the generally used backward analysis (see [4]). Then a linear system

$$B\varepsilon = \eta \tag{1}$$

with respect to the vector of equivalent perturbations ε is derived, and the solution of this system gives a first order approximation of the equivalent perturbations. Here matrix B consists of the Frechet-derivatives of all the operations and of elements which are equal either to 0 or to -1. η is the vector of all local absolute round-off errors. Giving values to the equivalent perturbations of the output data we can estimate successively, level by level (see [2], [3]), all the other equivalent perturbations. We are interested in the equivalent perturbations of the input data which are the results of the backward analysis.

The estimates of backward analysis can be written in a simple analytical form, while the estimates of forward analysis depend strongly on intermediate results. Besides, backward analysis needs much less operations when the estimates are defined numerically.

2. Description of the LU-decomposition. The algorithm is described in [1]. Let us consider the system

$$Ax = f \tag{2}$$

with

$$A = \begin{bmatrix} c_1 & d_1 & e_1 & & & & \\ b_2 & c_2 & d_2 & e_2 & & & \\ a_3 & b_3 & c_3 & d_3 & e_3 & & \\ & & & \cdots & & & \\ & & & a_{n-1} & b_{n-1} & c_{n-1} & d_{n-1} \\ & & & & a_n & b_n & c_n \end{bmatrix}, \quad f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_n \end{bmatrix}.$$

We look for a solution of the following kind:

$$\begin{aligned} &x_i = \alpha_i x_{i+1} + \beta_i x_{i+2} + \gamma_i, \quad i = 1, \ldots, n-2 \\ &x_{n-1} = \alpha_{n-1} x_n + \gamma_{n-1}, \\ &x_n = \gamma_n. \end{aligned} \tag{3}$$

Let us note that $\alpha_1, \beta_1, \gamma_1$ can be derived from the first equation of system (2), and then using the representation (3) for x_{i-2}, x_{i-1} we get the coefficients $\alpha_i, \beta_i, \gamma_i$ from the i-th equation as follows:

$$\begin{aligned} \Delta_1 &= c_1, \ \alpha_1 = -d_1/\Delta_1, \ \beta_1 = -e_1/\Delta_1, \ \gamma_1 = f_1/\Delta_1, \\ \Delta_2 &= c_2 + b_2\alpha_1, \ \alpha_2 = -(d_2 + b_2\beta_1)/\Delta_2, \ \beta_2 = -e_2/\Delta_2, \\ \gamma_2 &= (f_2 - b_2\gamma_1)/\Delta_2, \\ \Delta_i &= c_i + (a_i\alpha_{i-2} + b_i)\alpha_{i-1} + a_i\beta_{i-2}, \\ \alpha_i &= -[d_i + (a_i\alpha_{i-2} + b_i)\beta_{i-1}]/\Delta_i, \\ \beta_i &= -e_i/\Delta_i, \\ \gamma_i &= [f_i - (a_i\alpha_{i-2} + b_i)\gamma_{i-1} - a_i\gamma_{i-2}]/\Delta_i, \\ i &= 1, \ldots, n, \end{aligned} \tag{4,5}$$

where $e_{n-1} = e_n = d_n = 0$. Equalities (4) are called forward elimination, and equalities (3)-back substitution.. Actually, (3) and (4,5) realize the following decomposition $A = LU$, where

$$L = \begin{bmatrix} \Delta_1 & & & & \\ b_2 & \Delta_2 & & 0 & \\ a_3 & a_3\alpha_1 + b_3 & \Delta_3 & & \\ & \ddots & \ddots & \ddots & \\ & 0 & & & \\ & & a_n & a_n\alpha_{n-2} + b_n & \Delta_n \end{bmatrix}, \quad U = \begin{bmatrix} 1 & -\alpha_1 & -\beta_1 & & 0 \\ & 1 & -\alpha_2 & -\beta_2 & \\ & & \ddots & \ddots & \\ & 0 & & 1 & -\alpha_{n-1} \\ & & & & 1 \end{bmatrix}.$$

From (4,5) we obtain the triangular system

$$U_x = \gamma, \quad \gamma = L^{-1} f, \tag{6}$$

where $\gamma = (\gamma_1, \ldots, \gamma_n)^T$, and then the recurrence relations (3) produce the solution x.

The round-off error analysis is done under the assumptions that matrix A is diagonally dominant, i.e.

$$|c_i| \geq |a_i| + |b_i| + |d_i| + |e_i|, \quad i = 1, \ldots, n, \tag{7}$$

for $a_1 = b_1 = a_2 = e_{n-1} = e_n = d_{n-1} = 0$, and that at least for one i the inequality is strict. Under these assumptions it can be shown that the algorithm is correct (see [1]) and that the following estimate is valid:

$$|\alpha_i| + |\beta_i| \leq 1, \quad i = 1, \ldots, n. \tag{8}$$

3. Backward analysis of the back substitution. We shall do the backward analysis of the forward elimination and the back substitution separately. Let us consider the back substitution at first. The dependence graph of this part of the algorithm is given in fig. 1, where $q_i = (\alpha_i, \beta_i, \gamma_i)$. In each vertex only one term of the recurrence relation (3) is computed. The vectors q_i are inputs for the back substitution. Now using the method described in [3] we see that matrix B from (1) has the following structure:

$$B = \left[\begin{array}{ccccccc:ccccc}
\tilde{x}_n & 1 & & & & 0 & & \tilde{\alpha}_{n-1} & -1 & & & \\
 & & & & & & & & & & 0 & \\
 & & & \tilde{x}_{n-1} & \tilde{x}_n & 1 & & \tilde{\beta}_{n-2} & \tilde{\alpha}_{n-2} & -1 & & \\
 & \cdots\cdots & & \cdots & & & & & \cdots\cdots & \cdots & & \\
 & 0 & & & & & & 0 & & & & \\
 & & & & & \tilde{x}_2 & \tilde{x}_3 \quad 1 & & & \tilde{\beta}_1 & \tilde{\alpha}_1 & -1
\end{array}\right] \tag{9}$$

The wave denotes that the elements are computed with round-off errors. The size of matrix B is $(n-1) \times (4n-4)$ and it has a full rank. System (1) has a set of solutions and we have a choice.

Using floating-point arithmetic operations we assume that $f\bar{l}(x * y) = (x * y)(1+\rho)$, for $* \in \{+, -, \times, /\}$, where $|\rho| \leq 0.5p^{-t+1}$, p is the radix of the number system, and t is the number of mantissa digits (see [4]).

Further on the lower indices of ε and η denote the corresponding equivalent perturbations and absolute round-off errors. Then neglecting terms of second order in p^{-t+1} simple round-off analysis gives that

$$\eta_{x_{n-1}} = \tilde{\alpha}_{n-1}\tilde{x}_n(\rho_1^{(n-1)} + \rho_2^{(n-1)}) + \tilde{\gamma}_{n-1}\rho_2^{(n-1)},$$

$$\eta_{x_i} = \tilde{\alpha}_i\tilde{x}_{i+1}(\rho_1^{(i)} + \rho_3^{(i)} + \rho_4^{(i)}) + \tilde{\beta}_i\tilde{x}_{i+1}(\rho_2^{(i)} + \rho_3^{(i)} + \rho_4^{(i)}) + \tilde{\gamma}_i\rho_4^{(i)},$$
$$|\rho_j^{(i)}| \le 0.5p^{-t+1}, \quad i = n-2,\dots 1, \quad j = 1,\dots,4.$$

Now we choose the following solution of system (1) with matrix (9) :

$$\begin{aligned}
&\varepsilon_{x_i} = 0, \quad i = n,\dots,1,\\
&\varepsilon_{\alpha_{n-1}} = \tilde{\alpha}_{n-1}(\rho_1^{(n-1)} + \rho_2^{(n-1)}), \quad \varepsilon_{\gamma_{n-1}} = \tilde{\gamma}_{n-1}\rho_2^{(n-1)},\\
&\varepsilon_{\alpha_i} = \tilde{\alpha}_i(\rho_1^{(i)} + \rho_3^{(i)} + \rho_4^{(i)}),\\
&\varepsilon_{\beta_i} = \tilde{\beta}_i(\rho_2^{(i)} + \rho_3^{(i)} + \rho_4^{(i)}),\\
&\varepsilon_{\gamma_i} = \tilde{\gamma}_i\rho_4^{(i)}, |\rho_j^{(i)}| \le 0.5p^{-t+1}, \quad i = n-1,\dots,1, \quad j = 1,\dots,4.
\end{aligned} \tag{10}$$

Besides, let us have $\varepsilon_{\gamma_n} = 0$. From (10) we obtain the estimates:

$$\begin{aligned}
&\varepsilon_{\gamma_n} = 0,\\
&|\varepsilon_{\alpha_{n-1}}| \le |\tilde{\alpha}_{n-1}|p^{-t+1}, |\varepsilon_{\gamma_{n-1}}| \le 0.5p^{-t+1},\\
&|\varepsilon_{\alpha_i}| \le 1.5|\tilde{\alpha}_i|p^{-t+1},\\
&|\varepsilon_{\beta_i}| \le 1.5|\tilde{\beta}_i|p^{-t+1},\\
&|\varepsilon_{\gamma_i}| \le 0.5|\tilde{\gamma}_i|p^{-t+1}, \quad i = n-2,\dots,1.
\end{aligned} \tag{11}$$

So, the backward analysis of the back substitution gives very good estimates of both absolute and relative equivalent perturbations.

4. Backward analysis of the forward elimination. The dependence graph of this part of the algorithm is given in fig. 2. There operations (4) are placed in the first two vertices, and the operation (5) is placed in every other vertex in the graph, where the vectors $r_i = (a_i, b_i, d_i, e_i, c_i, f_i)^T$ are the inputs. In this case matrix B has the following structure:

$$B = \begin{bmatrix}
H_1 & -I & & & & & & & & \\
 & & & & & & & 0 & & \\
 & & G_2 & H_2 & -I & & & & & \\
 & & F_3 & 0 & G_3 & H_3 & -I & & & \\
 & & & \cdots\cdots\cdots & & & & & & \\
 & & 0 & & & & & & & \\
 & & & & & F_n & 0 & G_n & H_n & -I
\end{bmatrix},$$

where F_i, G_i, H_i are Frechet-derivatives of the operations in every vertex of the graph with respect to the vectors $(\tilde{\alpha}_{i-2}, \tilde{\beta}_{i-2}, \tilde{\gamma}_{i-2})^T$, $(\tilde{\alpha}_{i-1}, \tilde{\beta}_{i-1}, \tilde{\gamma}_{i-1})^T$ and r_i. Blocks F_i, G_i, H_i are given below:

$$F_i = \frac{-a_i}{\tilde{\Delta}_i}\begin{bmatrix}
\tilde{\beta}_{i-1} - \tilde{\alpha}_{i-1}\tilde{\alpha}_i & -\tilde{\alpha}_i & 0\\
-\tilde{\alpha}_{i-1}\tilde{\beta}_i & -\tilde{\beta}_i & 0\\
\tilde{\gamma}_{i-1} + \tilde{\alpha}_{i-1}\tilde{\gamma}_i & -\tilde{\gamma}_i & -1
\end{bmatrix}, \tag{12}$$

(13)
$$G_i = \frac{a_i\tilde{\alpha}_{i-2} + b_i}{\tilde{\Delta}_i}\begin{bmatrix} -\tilde{\alpha}_i & 1 & 0 \\ -\tilde{\beta}_i & 0 & 0 \\ -\tilde{\gamma}_i & 0 & -1 \end{bmatrix},$$

$$H_i = \begin{bmatrix} \frac{\partial\tilde{\alpha}_i}{\partial a_i} & \frac{\partial\tilde{\alpha}_i}{\partial b_i} & \frac{1}{\tilde{\Delta}_i} & 0 & \frac{-\tilde{\alpha}_i}{\tilde{\Delta}_j} & 0 \\ \frac{\partial\tilde{\beta}_i}{\partial a_i} & \frac{\partial\tilde{\beta}_i}{\partial b_i} & 0 & \frac{-1}{\tilde{\Delta}_i} & \frac{-\tilde{\beta}_i}{\tilde{\Delta}_j} & 0 \\ \frac{\partial\tilde{\gamma}_i}{\partial a_i} & \frac{\partial\tilde{\gamma}_i}{\partial b_i} & 0 & 0 & \frac{-\tilde{\gamma}_i}{\tilde{\Delta}_i} & \frac{1}{\tilde{\Delta}_i} \end{bmatrix}.$$

Here we assume that $\tilde{\Delta} \neq 0$, $i = 2, \ldots, n$. The derivatives with respect to a_i and b_i are not necessary in the further investigation, so, they are not written explicitly. The equivalent perturbations $\varepsilon_{\alpha_i}, \varepsilon_{\beta_i}, \varepsilon_{\gamma_i}$ are already defined in Section 2. For this reason and from the structure of matrix B in this section it is clear that we have to solve a system with the block diagonal matrix diag $\{H_i\}_{i=1}^n$ in order to obtain the equivalent perturbations of the vectors r_i. Here we consider only the i-th block equation. It looks as follows:

(14)
$$H_i\varepsilon_{r_i} = \eta_{q_i} - F_i\varepsilon_{q_{i-2}} - G_i\varepsilon_{q_{i-1}} + \varepsilon_{q_i}.$$

Neglecting terms of second order in p^{-t+1} simple round-off error analysis gives the estimates of $\eta_{q_i} = (\eta_{\alpha_i}, \eta_{\beta_i}, \eta_{\gamma_i})^T$:

(15)
$$\begin{aligned} |\eta_{\alpha_i}| &\leq (|d_i| + 2.5|a_i\tilde{\alpha}_{i-2}\tilde{\beta}_{i-1}| + 2|b_i\tilde{\beta}_{i-1}|)|\tilde{\Delta}_i^{-1}|p^{-t+1} + |\tilde{\alpha}_i||\tilde{\Delta}_i^{-1}||\eta_{\Delta_i}|, \\ |\eta_{\beta_i}| &\leq 0.5|e_i||\tilde{\Delta}_i^{-1}|p^{-t+1} + |\tilde{\beta}_i||\tilde{\Delta}_i^{-1}||\eta_{\Delta_i}|, \\ |\eta_{\gamma_i}| &\leq (1.5|f_i| + 3|a_i\tilde{\gamma}_{i-1}\tilde{\alpha}_{i-2}| + 2.5|b_i\tilde{\gamma}_{i-1}| \\ &\quad + 1.5|a_i\tilde{\gamma}_{i-2}|)|\tilde{\Delta}_i^{-1}|p^{-t+1} + |\tilde{\gamma}_i||\tilde{\Delta}_i^{-1}||\eta_{\Delta_i}|, \end{aligned}$$

where

$$|\eta_{\Delta_i}| \leq (|c_i| + 2.5|a_i\tilde{\alpha}_{i-1}\tilde{\alpha}_{i-2}| + 2|b_i\tilde{\alpha}_{i-1}| + |a_i\tilde{\beta}_{i-2}|)p^{-t+1}.$$

System (14) has a set of solutions. Let us choose $\varepsilon_{a_i} = \varepsilon_{b_i} = \varepsilon_{c_i} = 0$. Then the rest of the unknown $\varepsilon_{r_I}^* = (\varepsilon_{d_i}, \varepsilon_{e_i}, \varepsilon_{f_i})^T$ are defined uniquely:

$$\varepsilon_{r_i}^* = \tilde{\Delta}_i(\eta_{q_i} - F_i\varepsilon_{q_{i-2}} - G_i\varepsilon_{q_{i-1}} + \varepsilon_{q_i}).$$

In all the following estimates neglecting terms of second order in p^{-t+1} we can consider that

(16)
$$|\tilde{\alpha}_i + \tilde{\beta}_i| \leq 1.$$

Now from (11), (12), (13) and (15) after some computations one can obtain the following estimates:

$$\|\varepsilon_A\|_\infty \le \max_i(5|c_i| + |d_i| + 14|a_i| + 10|b_i| + 0.5|e_i|)p^{-t+1} \le \le 9.5\|A\|_\infty p^{-t+1}, \tag{17}$$

$$\|\varepsilon_f\|_\infty \le \max_i[1.5|f_i| + (13|a_i| + 7|b_i| + 1.5|c_i|)\|\tilde{\gamma}\|_\infty]p^{-t+1} \le \le (1.5\|f\|_\infty + 7.25\|A\|_\infty\|\tilde{\gamma}\|_\infty)p^{-t+1}. \tag{18}$$

The last estimate depends on the intermediate data $\tilde{\gamma}$. Two other estimates follow from (18) and (6) depending only on input or output data:

$$\|\varepsilon_f\|_\infty \le (2\|f\|_\infty + 11.5\|A\|_\infty\|\tilde{x}\|_\infty)p^{-t+1},$$

$$\|\varepsilon_f\|_\infty \le (2 + 11.5\|A\|_\infty\|A^{-1}\|_\infty)\|f\|_\infty p^{-t+1}.$$

Here we use the fact that $\|U\|_\infty \le 2$ and $L^{-1} = UA^{-1}$. The estimates thus obtained depend only on the condition of A and do not depend explicitly on n. This shows that the algorithm is stable and backward analysis depends only on the condition of problem (2).

In the next section two other estimates are used. They follow from (17) and (18):

$$\|\varepsilon_A\|_\infty \le (5\max_i|c_i| + \max_i|d_i| + 14\max_i|a_i| + 10\max_i|b_i| + 0.5\max_i|e_i|)p^{-t+1}, \tag{19}$$

$$\|\varepsilon_f\|_\infty \le [1.5\max_i|f_i| + (13\max_i|a_i| + 7\max_i|b_i| + 1.5\max_i|c_i|)\|\tilde{\gamma}\|_\infty]p^{-t+1}. \tag{20}$$

Let us note that forward analysis can be obtained from system (1) using the representations of the blocks F_i, G_i, H_i, but it depends on the quantities $|a_i|/|\Delta_i|$, $|a_i\alpha_{i-2} + b_i|/|\Delta_i|$, which cannot be estimated analytically so easily. Besides, backward analysis uses $(7n-6)$ comparisons and 16 multiplications and additions in (19) and (20), while forward analysis would use $0(n)$ arithmetic operations.

5. Numerical results. The experiments are realized on PC/AT where $p^{-t+1} \approx 10^{-7}$. The algorithm is tested with matrices of order $n = 20, 50, 100, 200, 500, 1000, 2000$, the coefficients of which are given in Table 1. The systems with matrices from $M2(n)$ are solved for $\rho = 0.001, 0.12, 0.25, 0.5, 1, 2, 4, 100$. The right part f is chosen so, that the exact solution is $x = (1, \ldots, 1)^T$ in all examples. The estimates of the equivalent perturbations $EP = \|\varepsilon_A\|_\infty + \|\varepsilon_f\|_\infty$ from (19) and (20) and the quantity $ERR = \|x - \tilde{x}\|_\infty$ are compared in all the tests, where $\tilde{x}$ is the solution of (2) with round-off errors and x is the exact solution.

For the class $M1(n)$ we have $EP \le 6 \times 10^{-6}$ for all n and $ERR \le 10^{-6}, 10^{-6}$, 1.9×10^{-6}, $1.3 \times 10^{-5}, 6.9 \times 10^{-5}$, 7.8×10^{-5}, 2.9×10^{-4} for the corresponding n. Small EP shows that the algorithm is stable. The norm $\|A^{-1}\|_\infty$ is growing with n.

	a_i	b_i	c_i	d_i	e_i
$M1(n)$	-1	-1	4	-1	-1
$M2(n)$	$-\rho$	$-\rho$	$1+4\rho$	$-\rho$	$-\rho$
$M3(5)$	-1	-1	$2, i=1$ $102, i=2$ $3+10^{2i-2}, i=3,4$ $2, i=5$	-10^{2i-2}	-1
$M4(10)$	-1	-1	$2, i=1$ $12, i=2$ $3+10^{i-1}, i=3,\ldots,9$ $2, i=10$	-10^{i-1}	-1

Table 1

Table 2 shows the results for the matrices of $M2(n)$. The quantities EP and ERR rarely change with the growth of n.

ρ	EP	ERR
0.001	5.72×10^{-7}	1.19×10^{-7}
0.12	1.29×10^{-6}	1.19×10^{-7}
0.25	2.06×10^{-6}	1.19×10^{-7}
0.5	3.56×10^{-6}	1.19×10^{-7}
1	6.55×10^{-6}	1.19×10^{-7}
2	1.25×10^{-5}	2.38×10^{-7}
4	2.45×10^{-5}	1.19×10^{-6}

Table 2

The results for $M2(n)$ when $\rho = 100$ are given separately in Table 3 because ERR changes for different n. Although these matrices are ill-conditioned ($\|A^{-1}\|_\infty \geq 10^6$) Table 3 shows that the equivalent perturbations describe the behavior of the round-off error quite well.

n	EP	ERR
20	6×10^{-4}	1.19×10^{-6}
50	6×10^{-4}	2.62×10^{-6}
100	6×10^{-4}	2.62×10^{-6}
200	6×10^{-4}	1.67×10^{-5}
500	6×10^{-4}	2.34×10^{-5}
1000	6×10^{-4}	2.34×10^{-5}
2000	6×10^{-4}	2.34×10^{-5}

Table 3

Finally for the matrix $M3(5)$ we have that $EP \leq 5.42 \times 10^{-1}, ERR \leq 3.4 \times 10^{-2}$, and for the matrix $M4(10)$ we have that $EP \leq 73, ERR \leq 3$. The equivalent perturbations describe the real situation quite well again. The last two examples also

show that although matrices $M3(5)$ and $M4(10)$ are diagonally dominant and the diagonal dominance is strict for one row of these matrices, the result is far away from the exact solution x. The explanation is that for these matrices the coefficients α_i are approaching 1, the coefficients β_i are approaching 0, because the elements d_i are growing very fast. For this reason we have $\Delta_n \approx 0$ ($\Delta_5 = 1.93 \times 10^{-6}$, for $M3(5)$, $\Delta_{10} = 1.92 \times 10^{-8}$, for $M4(10)$), and $\gamma_n = x_n$ is computed with big round-off error.

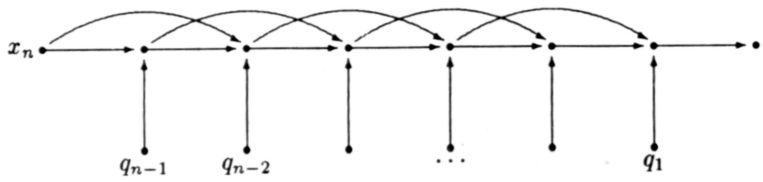

Fig. 1.
Dependence graph of the back substitution

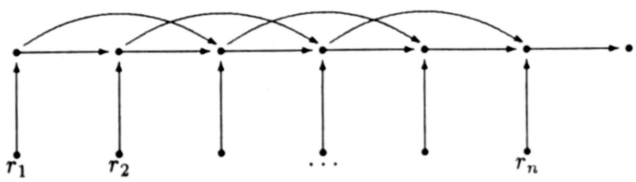

Fig. 2.
Dependence graph of the forward elimination

REFERENCES

[1] SAMARSKIY, A.A., E.S. NIKOLAEV. Methods for solving grid equations, Moscow, 1978 (in Russian).

[2] VOEVODIN, V.V. Mathematical models and methods in parallel processing, Moscow, 1986 (in Russian).

[3] VOEVODIN, V.V., P.Y. YALAMOV. A new method of round-off error estimation. Proc. Workshop on Parallel and Distributed Processing, March 27-29, Elsevier, Amsterdam, 1990.

[4] J.H. WILKINSON, J.H. The algebraic eigenvalue problem, Clarendon Press, Oxford, 1965.

Department of Mathematics
Technical University
7017 Russe
BULGARIA

Received 4.02.1992