Opl Res. Q., Pergamon Press 1977. Vol. 28, ii, pp. 641 to 651. Printed in Great Britain

Planning Emergency Ambulance Services*

KENNETH N. GROOM

NHS Operational Research Group, Reading

An evaluation model of an emergency ambulance service has been prepared to estimate the response time distributions given by various arrangements of vehicles and operating conditions. The model combines the *range* of vehicles with their *availability*. Range is a geographical variable determined by ambulance locations and travel speeds. Availability, which is governed by incident frequency and the time required to replace vehicles attending incidents, is obtained from standard queuing theory.

The model is used in a dialogue with ambulance service planners to evaluate the available options leading to the formulation of a plan for emergency cover in the area. An application is described for the West Glamorgan Area Health Authority.

INTRODUCTION

THIS paper describes the use of a general model[1] developed by the National Health Service Operational Research Group. The model has been used in several applications (Appendix 1) but in this paper a single application, for the West Glamorgan Area Health Authority,[2] is used as an example. The work was concerned with the problem of relating the provision of resources to the service given to patients by the ambulance service. Knowledge of this relation is essential for any consideration of fundamental planning decisions in this emergency service.

Convenient measures of performance were obtained from the standards recommended for the monitoring system described in a health service circular.[3] These were determined by the Orcon Group[4] as levels that most authorities could expect to achieve. For emergencies the standards recommended were as follows:

—*activation time*, the interval between notification of an incident and the despatch of a fully equipped and crewed vehicle, should be 3 min or less for 95% of all calls;

—*response time*, the interval between notification of an incident and the arrival at scene, should be 8 min or less for 50% of all calls, and 20 min or less for 95% of all calls (for metropolitan services the standard times are 7 and 14 min for 50 and 95% respectively).

Good performance against these standards is clearly necessary although not sufficient to provide an adequate service, other less readily measurable

* This is a revised version of a paper presented at the O.R. Society conference, Swansea, 1976.

elements (skill, courtesy, etc.) are evidently important. However, activation and response times do provide a useful yardstick for both monitoring and planning emergency cover. In our studies we have taken the expected cumulative distribution of response time as the measure of cover. Thus the cover within, say, 8 min is the proportion of calls that can be expected to have a response time of 8 min or less.

EMERGENCY COVER MODEL

To calculate cover, information is required on three topics:

—*geography*, where are the roads and how good are they?
—*incidents*, what frequency and where?
—*organisation*, how is the service organised?

Geography. The area is represented by a set of nodes that are interconnected by a road network. Nodes exist for population centres, major road junctions, hospitals, existing and potential ambulance station locations. Figure 1 is a sketch map indicating the frame used for data collection and analysis in West Glamorgan. Travel times between adjacent nodes for emergency ambulances are estimated and used with the aid of a shortest path algorithm[5] to build up a complete matrix of travel times between all pairs of nodes.

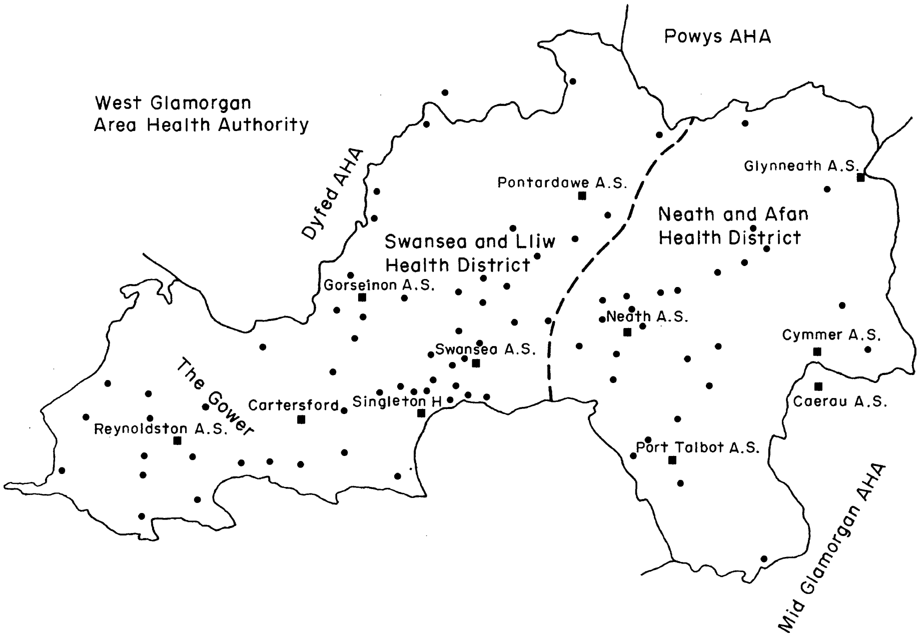


Fig. 1

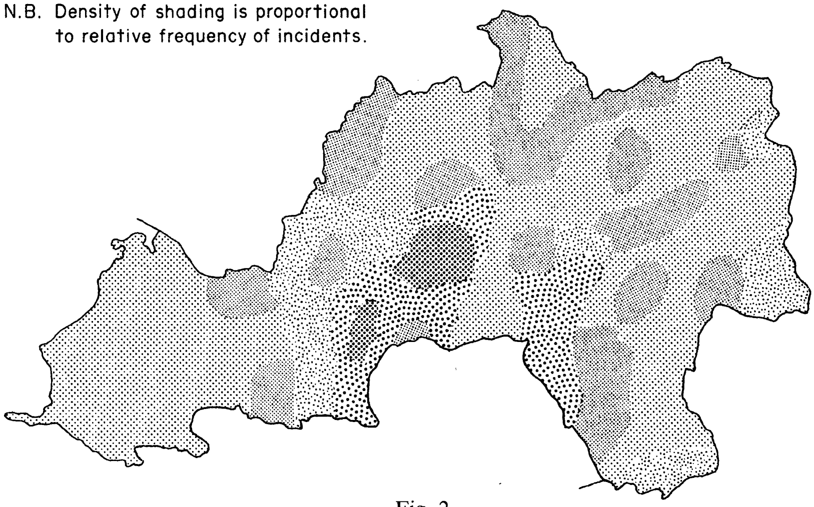

N.B. Density of shading is proportional
to relative frequency of incidents.

Fig. 2

Incidents. The geographical distribution of emergency calls, and their frequency is assembled from past records for each shift—weekday/weekend, summer and winter. The map shown as Fig. 2 indicates the average distribution of incidents for West Glamorgan.

Organisation. The procedure for dealing with emergencies may vary between day, night or weekends. The key factors that define the operating conditions are:

(i) whether emergency cover can be replenished by ambulances on non-emergency duty—i.e. by operating a single-tier service.

(ii) whether the service is operated with back-up—i.e. by relocating stand-by ambulances to busy sites to restore cover after an ambulance has been despatched,

(iii) what value of activation time is appropriate to the control system being used.

The model uses this information to calculate the two basic ingredients of cover: *range* and *availability.*

range, $R_r(t)$ = the proportion of potential emergencies that can be reached within time t by any of the r ambulances.

availability, P_r = the proportion of time that r ambulances are available to respond to emergency calls.

Range and availability are combined to estimate $C_n(t)$, the cover given by n emergency ambulances in the equation:

$$C_n(t) = \sum_{r=0}^{n} P_r R_r(t).$$

The methods used in the calculation of range and availability are indicated in Appendix 2. The model is incorporated into several computer programs but these do not comprise a package guaranteed to give the 'right' answer, rather they are tools to be used with some skill and understanding. In each application the tools are used to explore the sensitivity to cover of each of many variables—not least the method's own assumptions. Once this has been done the tools can be further used to answer questions in the form "What will happen if...?"

USING THE MODEL

Predicting the present

The first question posed in each application is 'What will happen if no changes are made?' The results of this evaluation provide us with a calibration check, a verification of the basic data and initiates the dialogue with the authority that forms an essential part of all application studies. In West Glamorgan during the day shift the existing arrangement had 10 vehicles located at:

Swansea	3 vehicles
Port Talbot	1 vehicle
Neath	1 vehicle
Glynneath	1 vehicle
Pontardawe	1 vehicle
Gorseinon	1 vehicle
Reynoldston	1 vehicle
Cymmer	1 vehicle

They operated as a single tier together with twenty nonemergency vehicles and the model showed they would give 45 and 99% of response times within 8 and 20 minutes respectively, i.e. meeting the 20 minute standard but failing at the 8 min standard. Figure 3 shows the range of the 10 vehicles.

Evaluating future options

Having established the credentials of the model through its prediction of the present, the next part of the dialogue was concerned with finding an arrangement that would meet both the standards. The arrangement that meets the standards with least vehicles uses only six, located one each at:

Swansea
Singleton Hospital
Port Talbot
Neath
Glynneath
Pontardawe

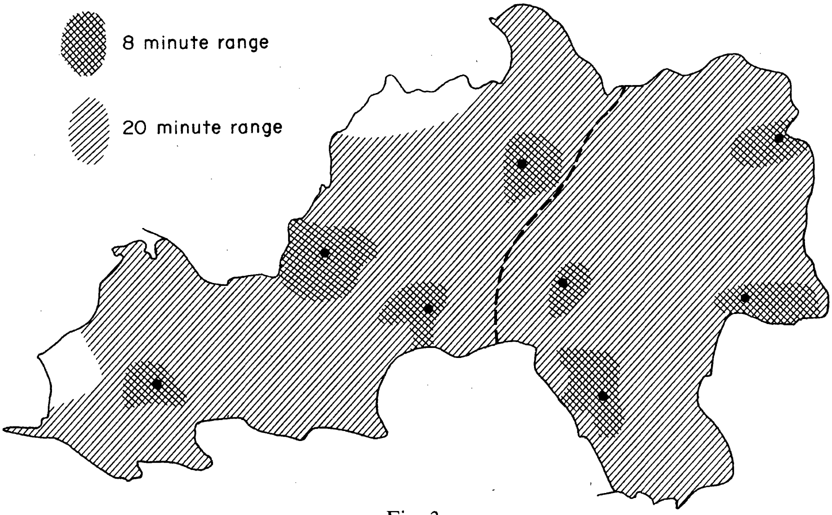


Fig. 3

These could meet 55 and 97% of responses within 8 and 20 min. However, this option did raise a series of questions in the dialogue that were considered in further evaluations.

Balance of cover between health districts

Although the service is organised on an Area basis the authority wanted to avoid large differences between the level of service given in each of the two District Health Authorities. In fact the 6 vehicle arrangement would give levels of cover that are up to standard in each health district but the maximum expected response times were quite different (see Table 1).

For the Neath and Afan health district, a maximum response of 26 min did not seem unreasonable and anyway further evaluations showed it could be reduced by cross-border agreement whereby cover is given to the Cymmer area by the Mid-Glamorgan station at Caerau.

For the Swansea and Lliw health district, however, 42 min was thought unreasonable and no cross-border arrangements could help the Gower where the deficiency was felt. Analysis of the Gower as a sub-area revealed that

Table 1. Cover to each health district with six vehicle arrangement

	Swansea & Lliw	Neath & Afan
Cover in 8 min (%)	57	50
Cover in 20 min (%)	97	96
Maximum response (min)	42	26

an additional vehicle located at Catersford gave the best compromise between vehicle utilisation and reduced maximum response time (a site further west would give greater reductions in long response times but would be used very infrequently).

Integrated radio control. All the arrangements evaluated had assumed the benefits of an integrated radio control system that had not, at the time of the study, been installed. Since some period for settling down the new system was necessary, it was agreed that it would be prudent to operate with a second vehicle at Swansea, the need for which could be reviewed later.

Thus the arrangement finally recommended used 8 vehicles located at:

Swansea	2 vehicles
Singleton Hospital	1 vehicle
Port Talbot	1 vehicle
Neath	1 vehicle
Glynneath	1 vehicle
Pontardawe	1 vehicle
Cartersford	1 vehicle

The range of these is shown in Fig. 4 and Table 2 gives the performance that could be expected during weekday shifts.

Similar analyses were conducted for the night shift, for weekends and holidays, and the final recommendations were subjected to sensitivity analysis to ensure their robustness against variations in the basic data that might occur seasonally or in peak hours.

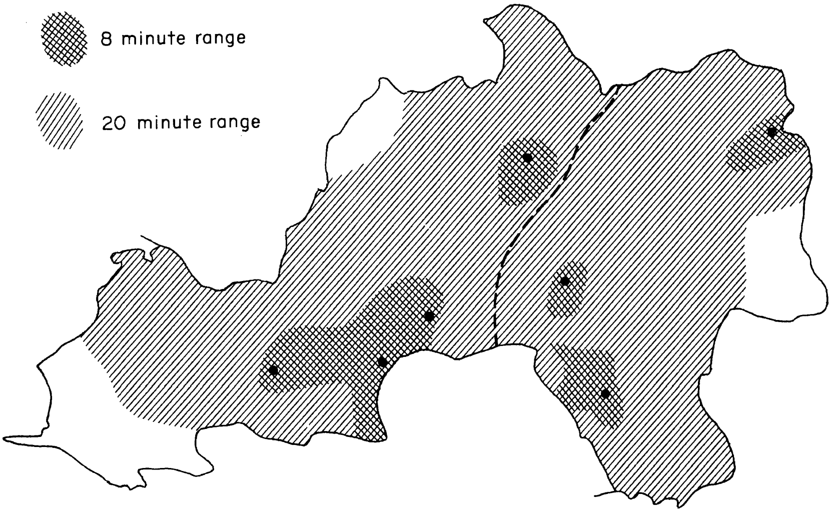


Fig. 4

TABLE 2. COVER GIVEN BY EIGHT VEHICLE ARRANGEMENT

	Whole AHA	Swansea HD	Neath HD
Cover in 8 min (%)	56	59	50
Cover in 20 min (%)	98	98	97
Maximum response (min)	32	32	26

ACKNOWLEDGEMENTS

I would like to express my appreciation of the ambulance officers and men who helped in this work and its many applications to the Department of Health and Social Security and the Welsh Office for their support and to the West Glamorgan AHA for permission to use their Service as an example. In words more appropriate to the location of the conference: "Mae'n pleser iawn gennyri ni i ddiolch y bobl hwn am eu help gyda'r waith."

APPENDIX 1. STUDIES USING THE METHOD

Complete (at 1st September, 1976)

Devon*
Dorset*
Dyfed
East Sussex
Greater Manchester—Metropolitan
Isle of Wight*
Mid-Glamorgan
Powys
Salop†
South Glamorgan
Suffolk
West Glamorgan
West Midlands—Metropolitan (Western Division)†

In progress

Clwyd†
Cornwall*
Essex†
Gwent
Gwynedd
Hampshire*
Kent
Lincolnshire
Oxfordshire

APPENDIX 2. THE CALCULATION OF RANGE AND AVAILABILITY

A full description of the model and the sensitivity analyses used to test its assumptions have been given elsewhere.[1] The purpose of this appendix is merely to indicate the approach that we adopted to calculate the two essential components of emergency cover and comment on assumptions implicit in the basic equation.

* Carried out by the National Coal Board O.R. Executive.

† Carried out by local Management Services staff with National Health Service Operational Research Group support.

Range

The range of exactly r ambulances is a spatial variable independent of emergency frequency. The range of r vehicles $R_r(t)$ is the proportion of potential emergencies that can be reached within time t by any of the r vehicles. This is simply the sum of the proportions of potential emergencies a_j at nodes that have a travel time from the r vehicle locations that is less than t. Thus:

$$R_r(t) = \sum a_j.$$

For all j such that travel time $d_{kj} \leqslant t$ where k is the location of the nearest vehicle to j.

Availability

A different method of calculation of availability has to be adopted for the different conditions of single and double tier operations.

In a *single tier* with N dual-purpose vehicles suitably equipped and crewed to deal with emergencies, n vehicles stand-by to deal with emergencies while the residual (N–n) vehicles attend to non-emergency duties. When an emergency call is received a stand-by ambulance is despatched to attend the call, and its place as stand-by vehicle is taken by the first vehicle to come available. This could be either a vehicle completing a non-emergency job or a vehicle completing an emergency job. If there is no deficiency in the number of stand-by vehicles when an emergency job is finished the vehicle coming free will take up non-emergency duties. There is thus an exchange of tasks between the vehicles on the fleet as emergencies arise and are dealt with.

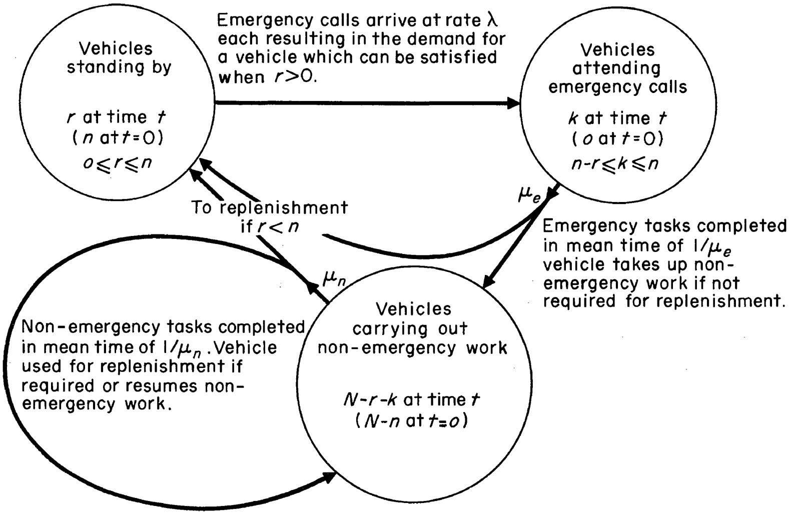

Vehicles standing by
r at time t
(n at t=0)
o≤r≤n
Emergency calls arrive at rate λ each resulting in the demand for a vehicle which can be satisfied when r>0.
Vehicles attending emergency calls
k at time t
(o at t=0)
n-r≤k≤n
To replenishment if r<n
μe
Emergency tasks completed in mean time of 1/μe vehicle takes up non-emergency work if not required for replenishment.
μn
Non-emergency tasks completed in mean time of 1/μn. Vehicle used for replenishment if required or resumes non-emergency work.
Vehicles carrying out non-emergency work
N-r-k at time t
(N-n at t=o)

Fig. A1

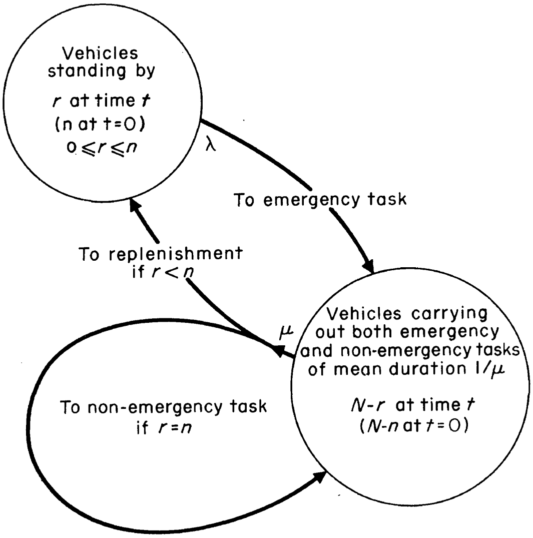

Vehicles standing by
r at time t
(n at t=0)
$0 \leqslant r \leqslant n$
λ
To emergency task
To replenishment if $r < n$
μ
Vehicles carrying out both emergency and non-emergency tasks of mean duration $1/\mu$
N-r at time t
(N-n at t=0)
To non-emergency task if $r = n$

Fig. A2

The flow of vehicles in this deployment/replenishment cycle is illustrated schematically in Fig. A1.

In practice it makes no difference to assume that the average service times for emergency and non-emergency tasks are equal. Thus the cycle can be simplified to that shown in Fig. A2.

Our task is to calculate the proportion of the time that there are n, $n - 1, \ldots, i, \ldots, 1, 0$ vehicles available on stand-by. These states occur when there are $0, 1, \ldots, n - i, \ldots, n - 1$, *n or more* unreplenished calls in the system. Calls arrive at rate λ and are replenished by any vehicle coming free—at an average rate μ per vehicle.

Following normal queuing theory procedure we obtain, for $\rho = \lambda/N\mu < 1$ (for $\lambda < N\mu$ there is no solution as would be expected when there are insufficient vehicles to meet the demands made on them), the required proportions:

$$p_n = q_0 = \left\{\sum_{i=0}^{n-1} \frac{(N-n)!}{(N-n+i)!}(N\rho)^i + \frac{(N-n)!}{N!}\frac{(N\rho)^n}{(1-\rho)}\right\}^{-1}$$

$$p_{n-i} = q_i = \frac{(N-n)!}{(N-n+i)!}(N\rho)^i p_n;\ 0 < i < n$$

$$p_o = \sum_{i=n}^{\infty} q_i = \frac{(N-n)!}{N!}\frac{(N\rho)^n}{(1-\rho)}p_n.$$

Double tier operations obviously apply when a service is organised in two tiers, but they also apply in single-tier services at night, during weekends

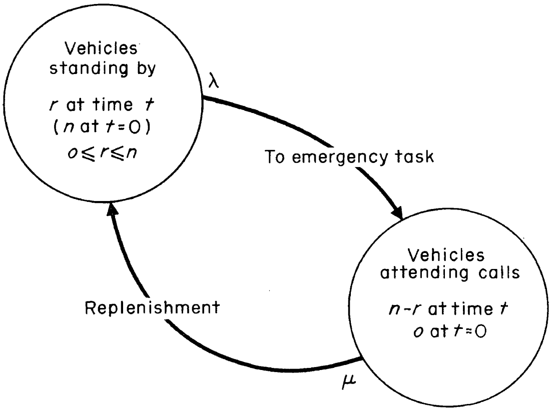


Fig. A3

and otherwise when there are no non-emergency vehicles in use. The flow of vehicles in an emergency only tier is illustrated in Fig. A3.

This system is simply an n-server queue and so the equations for availability are given by:

$$p_n = q_o = \left\{ \sum_{r=0}^{n-1} \frac{(n\rho)^r}{r!} + \frac{(n\rho)^n}{n!} \frac{1}{(1-\rho)} \right\}^{-1}$$

$$p_{n-i} = q_i = \frac{(n\rho)^i}{i!} p_n; \quad 0 < i < n; \quad \rho = \frac{\lambda}{n\mu} < 1$$

$$p_0 = \sum_{i=n}^{\infty} q_i = \sum_{i=n}^{\infty} \frac{n^n}{n!} \rho^i p_n = \frac{(n\rho)^n}{n!} \frac{p_n}{(1-\rho)}.$$

The basic equation

We have stated that cover is given by:

$$C_n(t) = \sum_{r=0}^{n} p_r R_r(t).$$

In making this statement we are implying that all states are covered in the expression. The implications of this differ between the models so are discussed separately.

The single-tier model assumed that when there is change of state (a vehicle despatched or replenished) that the remaining vehicles *relocate instantaneously* to stand-by positions appropriate for the different number of vehicles. We elaborate on these two key points below.

Relocate—the model assumes relocation of, in general, all vehicles on each change of state. In practice the sets of positions for $1, 2, \ldots, n$ vehicles are chosen so that each is a subset of the next. Thus, the relocation on a change of state would be of only one vehicle. This the model can and does reflect. However, in practice a controller would rarely relocate one vehicle as he would anticipate replenishment before the relocation would be effected. Thus, if sites were ordered so that the least used sites are vacated first, the model would assume a somewhat better cover than actually exists. Conversely, if the reverse order were taken, a worse than actual result would be indicated. This can be easily compensated for by using an alternative good site/bad site order. In practice this effect is not at all significant, the best and worst indications of cover at 8 and 20 min represent its middle value ± about 1%.

Instantaneous—the model assumes no period of transition between changes of state. Since the relocation effect can be compensated (or, as is more usual, ignored) it follows that assumptions concerning how long it takes to implement relocation are irrelevant.

The emergency-only tier models contain no relocation assumptions. The cover when r becomes $(r - 1)$ however will in general depend upon which of the r vehicles was deployed. There are two emergency-only tier models. The first is useful for areas with low traffic intensity and sums all call combinations for the first two terms in the cover series, provides a maximum likelihood estimator for the third and neglects all other terms. This model is valid for areas having low traffic intensities $(\lambda/n\mu)$. For higher traffic intensities more terms could be evaluated but we have adopted a different approach.

In the second model just one combination is evaluated, thus assuming calls arrive in specific order. Since the relative frequency of calls to each location is known: pessimistic, optimistic or other orders can be chosen at will. We know in practice that controllers do not relocate vehicles at every change of state. Nor do they allow the whole balance of cover to dissolve because of an unusual order of incidents. Careful choice of the orders tested allows a realistic evaluation reflecting realistic control and back-up arrangements.

REFERENCES

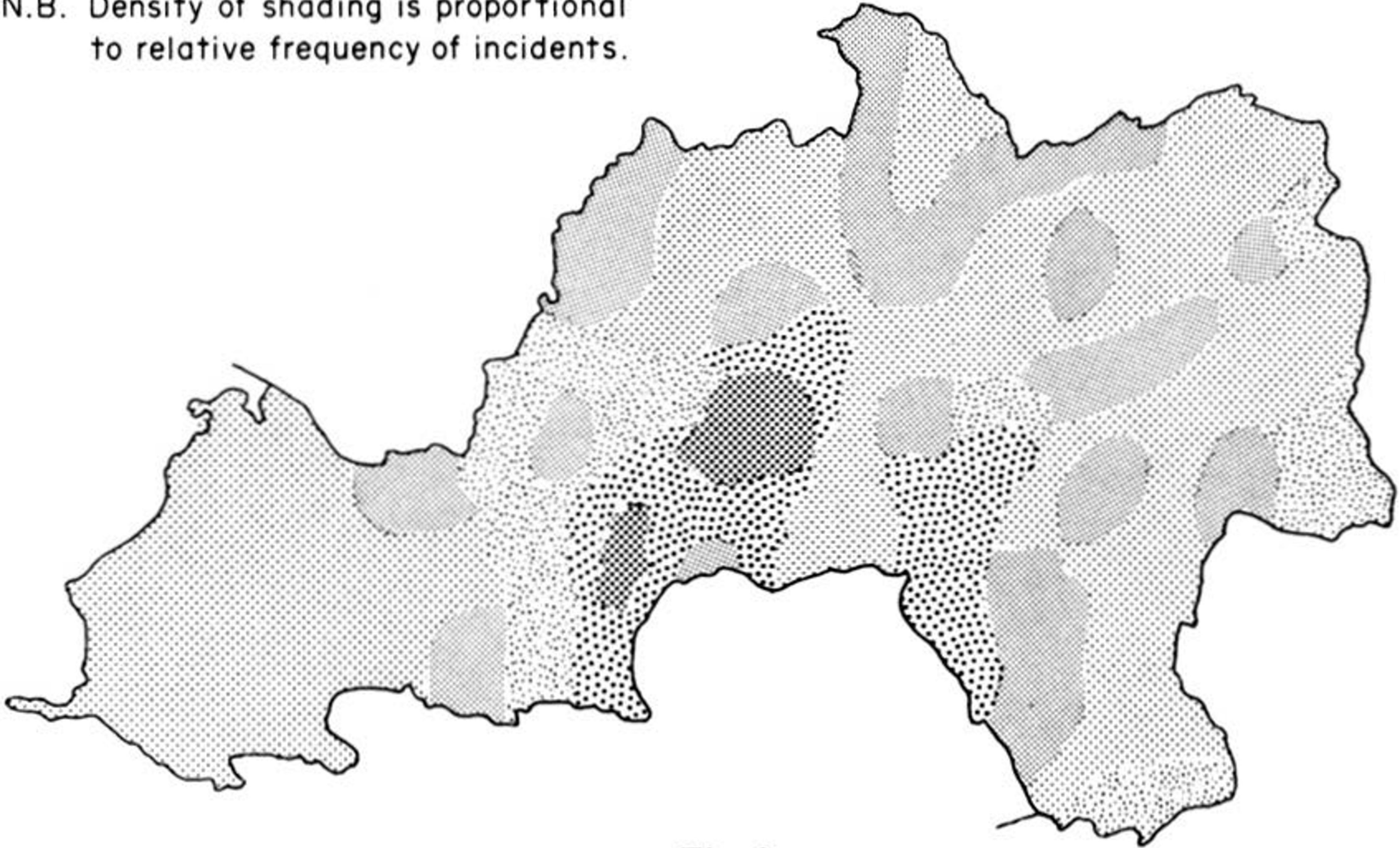
[1] K. N. Groom, *The Estimation of Emergency Ambulance Cover* NHS ORG Report 75/16, December, 1975.

[2] K. N. Groom, K. E. Holloway, and W. R. Mann (1975) *Planning Emergency Ambulance Cover in West Glamorgan*, NHS ORG Report 75/4, March.

[3] Organisation of Ambulance Services: Standards of Service and Incentive Bonus Schemes, Department of Health and Social Security/Welsh Office, HSC (IS) 67/WHSC (IS) 57, 1974.

[4] Ambulance Service Performance Standards and Measurements. Report to DHSS by Orcon Services, Cranfield Institute of Technology, 1974.

[5] R. W. Floyd (1962) Algorithm 97—Shortest path. *Commun. Ass. comput. Mach.* **5,** 345.



N.B. Density of shading is proportional to relative frequency of incidents.

Fig. 2