Theory and Methodology

Labour costs and queueing theory in retailing

J.B.G. Frenk, A.R. Thurik * and C.A. Bout

Econometric Institute, Erasmus University Rotterdam, 3000 DR Rotterdam, Netherlands

Received May 1989; revised February 1990

Abstract: In this paper approximation results for the M/G/*s* queueing model are used to derive an empirically verified shop type dependent non-homogeneous relation between labour volume and sales in retail trade. Moreover, we formulate the retailer's labour management as a formal minimization problem.

Keywords: Labour cost relation, queueing theory, approximations

1. Introduction

Non-homogeneous, linear labour costs relations are consistently found in empirical, cross-sectional studies of individual establishments in the retail trade (shops):

$$L_j = \beta_0 + \beta_1 Q_j, \tag{1}$$

where L is the annual labour volume per shop, Q the annual sales per shop, j the shop index and β_0, $\beta_1 > 0$. Relations of type (1) are found for shops belonging to a certain shop type. A shop type is defined as a class of shops which are similar with respect to service facilities, waiting time target and other shop characteristics such as assortment composition, degree of own production and type of organisation. Examples of shop types are chain supermarkets, independent butchers, cooperative clothes shops etc. See Nooteboom [8] and Thurik [15–17] for empirical evidence, and Nooteboom [9] for a survey of results.

The linear labour cost relation shows a certain threshold labour, β_0. This is intuitively appealing because there is no reason why the service capacity of a shop would approach zero if the use made of it by customers (measured in terms of sales) would: threshold labour consists of the minimum amount of labour if sales approach zero. In Nooteboom's model threshold labour equals the annual opening time of the shop. For instance, if weekly opening time is 60 hours and if, allowing for holidays etc. the shop is open during 45 weeks yearly, threshold labour is 2700 hours. This example describes the case in which a shop consists of one so-called (independently staffed) 'department'. If a shop consists of more than one 'department', threshold labour increases accordingly. Empirical tests confirm this a priori derived amount of threshold labour. See the references mentioned above where data are used from shop types with different opening times and with a different number of departments. Apart from differences in the level of threshold labour, further variation can be attributed to the use of part-time labour, quality of labour, depth of assortment, use of technology, mode of delivery, purchase volume per customer etc. The labour cost model has been estimated and

* Also at Research Institute for Small and Medium-Sized Business, Zoetermeer, Netherlands.

0377-2217/91/$03.50 © 1991 – Elsevier Science Publishers B.V. All rights reserved

corroborated using Dutch, German, French, British, American, Canadian and South-African data. See again the references mentioned above.

The results of the examples mentioned are used for forecasting and explaining labour productivity differences across individual shops. The results are also used explaining aggregate behaviour like productivity differences across shop types (see Thurik and Kleijweg [18]), productivity growth (see Nooteboom [10]) and the ousting of small business (see Nooteboom [11]).

The theoretical justification of (1) is given by Nooteboom [8,12]: queueing theory is used to analytically describe the labour requirement given demand characteristics and shop characteristics (service facilities). It is assumed that required labour is always available and that the fine tuning between labour capacity and labour requirement is accomplished using part-time labour.

Retailing does not produce a physical product that can be stocked or resold. Its product consists of a 'bundle of services' the capacity of which can be used by visiting customers. Hence, it is not straightforward that the concept of the traditional production function, where the level and combination of certain input factors determine sales, is relevant in retailing.

In retailing, production takes place when customers use the service capacity. Queueing theory is used to derive (1) in view of the stochastic nature of customers' arrivals. Moreover, labour costs are studied independently from other costs, such as occupancy costs. This is motivated by the definitional lack of substitution opportunities. If shop space would be used for substituting labour, the shop would end up in a different shop type because such a substitution would affect the characteristics of the marketing mix.

As opposed to Nooteboom we shall formulate the retailer's labour management problem as a formal minimization problem: the retailer aims to minimize labour costs with the restriction that the expected customer's waiting time, provided this customer has to wait, does not exceed a certain level. Furthermore, a rigorous derivation of the cost function (4) is given. (Compare (22) in Nooteboom [12].) The present more general model also leads to the non-homogeneous linear labour cost relation.

This concludes the introduction. In the next section a mathematical model describing the behaviour of the shop keeper is discussed and analyzed.

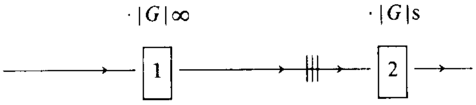

Fig. 1. A queueing model describing the flow of customers where node 2 represents the check-out point (cash registration) and node 1 the remaining activities

2. Model description and analysis

Before discussing the shopkeeper's objectives it is necessary to give a description of the flow of customers within a self-service shop. For that purpose we use a so-called tandem queue with infinite buffers between the nodes (cf. Figure 1)

The above model is a reflection of the fact that customers have to do all kinds of activities (node 1) before checking out (node 2). Hence, it consists of the following components:

(i) The arrival process of customers at node 1 is a Poisson process with parameter $\lambda > 0$.

(ii) Node 1 denotes a $\cdot/G/\infty$ queueing process, where the independent and identically distributed service time with finite first moment represent the sojourn times of the customers in the shop before entering the check-out waiting line.

(iii) Node 2 denotes the check-out point described by a $\cdot/G/s$ queueing process with FCFS discipline, service times S_i, $i \geq 1$, s servers (cashiers) and as arrival process the departure process from node 1.

Moreover, to keep the model analytically tractable we assume that

(iv) S_i, $i \geq 1$, is a sequence of independent and identically distributed random variables with $m_2 < \infty$ where $m_j := \mathscr{E}(S^j)$.

(v) The arrival process at node 1 and the service times at nodes 1 and 2 are independent.

Finally to ensure that the stochastic process $\{X_t(s), t \geq 0\}$ with

$$X_t(s) := \text{number of customers at node 2 at time } t \text{ whenever there are } s \text{ cashiers} \quad (2)$$

possesses a limiting distribution

$$p_j(s) := \lim_{t \uparrow \infty} P\{X_t(s) = j\}$$

we require that

(vi) $\lambda m_1 < s$.

This stochastic process will play an important role in the derivation of the empirically observed relation between labour volume and sales. By the above condition the stationary version $\boldsymbol{X}_t^*(s)$ of the stochastic process $\boldsymbol{X}_t(s)$ exists.

For a short proof of this result we observe (cf. [3]) that the arrival process at node 2 is a non-homogeneous Poisson process with intensity function $\lambda(t)$ equal to $\lambda B(t)$, where $B(t)$ denotes the distribution of the service times at node 1. Hence, if we condition on the value for $\boldsymbol{X}_T(s)$ with T fixed and arbitrary, the arrival process at node 2 can be stochastically bounded from below by a Poisson process with arrival rate $\lambda B(T)$ and from above with a Poisson process with arrival rate λ. This implies, since $\lambda m_1 < s$, that the corresponding bounding M/G/s models with initially $\boldsymbol{X}_T(s)$ customers in the system possess a limiting distribution (cf. [21]). Letting T tend to infinity these limiting distributions coincide and hence the result is proved. In the remainder we will now assume that the above model is in steady state or equilibrium at time 0.

Since we are only interested in the waiting time of a customer at node 2 (in steady state) the analysis of the waiting time at node 2 of the above tandem queue breaks down by the above remarks to the analysis of the waiting time in steady state of a M/G/s queue with arrival rate λ. Before introducing the approximation results available for the M/G/s queue we will derive a relation between labour volume and number of employed servants for a self-service shop. This relation together with the aimed service target of the shop keeper which can be realized by controlling the number of servants and hence depends on the stochastic process $\boldsymbol{X}_t^*(s)$ yields after some calculations the relation between labour volume and sales. By aggregating we finally obtain the desired result. In order to derive the relation between labour volume and employed servants we define (in the steady state) the following random variables:

$\boldsymbol{N}_t :=$ number of arriving customers at node 2 before time t, $t \le T$, with T denoting the annual opening time of a shop,

$\boldsymbol{W}_q(s) :=$ waiting time in the queue at node 2 of an arbitrary arriving customer whenever there are s cashiers,

$\boldsymbol{G}_c(s) :=$ total labour volume *available* for servicing customers during the annual opening time of a shop whenever there are s cashiers,

$\boldsymbol{G}_p(s) :=$ total labour volume to perform Pre-and Post Purchase Activities (PPA) *necessary* to operate the shop during one year whenever there are s cashiers; Nooteboom [12, p. 167] introduces PPA as activities that do not arise from direct dealings with customers such as physical distribution, breaking bulk, packaging, storage, price-labelling, display, own production, stock-taking, administration, cleaning, etc.

$\boldsymbol{G}(s) :=$ total labour volume *necessary* to operate the shop during one year whenever there are s cashiers (this includes also overtime!),

$\boldsymbol{Q}_i :=$ amount of sales of i-th entering (at node 2) customer,

$\boldsymbol{L}_i(s) :=$ the amount of time in $[0, T]$ cashier i, $i = 1, \ldots, s$ is idle,

$\boldsymbol{\rho}(s) :=$ fraction of total available labour volume during annual opening time of a shop *necessary* to service customers.

Clearly,

$$\boldsymbol{\rho}(s) = \frac{sT - \boldsymbol{I}(s)}{sT}, \tag{3}$$

with $\boldsymbol{I}(s) := \sum_{i=1}^{s} \boldsymbol{I}_i(s)$ representing the total amount of idle time in $[0, T]$, and

$$\boldsymbol{G}_c(s) = sT.$$

Moreover, by assumption (cf. [8,12]) $\boldsymbol{G}_p(s)$ is a linear function of total annual sales and hence

$$\boldsymbol{G}_p(s) = \gamma \sum_{i=1}^{\boldsymbol{N}_T} \boldsymbol{Q}_i.$$

By the randomness of the arrival process at node 2 cashiers are sometimes idle during opening hours and hence they can spend some fraction of their idle time to carry out PPA-activities. As an approximation of the total labour volume spent to perform PPA during opening hours we take $\mathscr{E}(\xi(\boldsymbol{\rho}(s)))\boldsymbol{I}(s)$ with $\xi : [0, 1] \to [0, 1]$ denoting some decreasing and continuous function satisfying $\xi(0) = 1$ and $\xi(1) = 0$. This implies that the total labour volume spent to perform PPA after opening hours during one year is roughly given by

$$\max\left(\boldsymbol{G}_{\mathrm{p}}(s)-\mathscr{E}(\xi(\rho(s)))\boldsymbol{I}(s),0\right).$$

Hence, as opposed to Nooteboom [12], total labour volume to perform PPA is regarded as a random variable. Since total labour volume consists of labour volume to service customers and labour volume due to PPA-activities, we obtain

$$\begin{aligned}\boldsymbol{G}(s)&\\ &=\boldsymbol{G}_{\mathrm{c}}(s)+\max\left(\boldsymbol{G}_{\mathrm{p}}(s)-\mathscr{E}(\xi(\rho(s)))\boldsymbol{I}(s),0\right)\\ &=sT+\max\left(\gamma\sum_{i=1}^{N_T}\boldsymbol{Q}_i-\mathscr{E}(\xi(\rho(s)))\boldsymbol{I}(s),0\right).\end{aligned}\tag{4}$$

Hence, the expected total labour volume for a homogeneous group of shop keepers having the same arrival rate of customers is given by

$$\begin{aligned}\mathscr{E}\boldsymbol{G}(s)&\\ &=sT+\mathscr{E}\left(\max\left(\gamma\sum_{i=1}^{N_T}\boldsymbol{Q}_i-\mathscr{E}(\xi(\rho(s)))\boldsymbol{I}(s),0\right)\right).\end{aligned}$$

Moreover, we assume that the shop keeper aims at a certain service target: he wishes to keep the expected waiting time of an arbitrary customer given that this customer has to wait below a certain level. Here we deviate from Nooteboom ([12]) in that we only consider customers who have to wait. Apart from the fact that it can be argued that this condition is more realistic it also makes the mathematical analysis more tractable especially for the case of negative exponentially distributed service times. Combining the above observations we obtain the following optimization problem:

$$\underset{s\in B}{\text{Minimize}}\quad\mathscr{E}\boldsymbol{G}(s),\tag{5}$$

with

$$B=\{s\colon \mathrm{CW}(s)\le D\},$$

$$\mathrm{CW}(s):=\mathscr{E}\left(\boldsymbol{W}_q(s)\mid\boldsymbol{W}_q(s)>0\right).$$

For this optimization problem, the following result can easily be proved.

Lemma 1. *The optimal solution of the above optimization problem equals* $\mathscr{E}\boldsymbol{G}(s_{\mathrm{OPT}})$ *where* $s_{\mathrm{OPT}}:=\min\{s\colon \mathrm{CW}(s)\le D\}$.

Proof. It is not difficult to verify that $\mathscr{E}\boldsymbol{G}(s)$ is increasing in s. This proves the result. □

In order to solve the above problem we need to compute s_{OPT} or equivalently to derive an analytical expression for $\mathrm{CW}(s)$. Before doing so, define for the M/G/s model

$$p_n:=\lim_{t\uparrow\infty}P\left\{\begin{matrix}\text{at time } t \text{ there are } n\\ \text{customers in the system}\end{matrix}\right\},\quad n\ge 0,$$

$$p_W:=\sum_{n=s}^{\infty}p_n,$$

$$\Omega:=\left(\sum_{k=0}^{s-1}\frac{(\lambda m_1)^k}{k!}+\frac{(\lambda m_1)^s}{s!(1-\rho)}\right)^{-1},$$

$$\rho:=\frac{\lambda m_1}{s}<1.$$

We write $p_n=p_n(\exp)$, $P_W=P_W(\exp)$, $\mathscr{E}\boldsymbol{W}_{\mathrm{q}}(s)=\mathscr{E}^{\exp}(\boldsymbol{W}_{\mathrm{q}}(s))$, $\mathrm{CW}(s)=\mathrm{CW}^{\exp}(s)$ when the service time is exponentially distributed with first moment m_1. In this case we have the explicit results (cf. [3, p.88]).

$$p_n(\exp)=\frac{(\lambda m_1)^n}{n!}\Omega,\quad 0\le n\le s,$$

$$P_W(\exp)=\frac{(\lambda m_1)^s}{s!(1-\rho)}\Omega,$$

$$\mathscr{E}^{\exp}\left(\boldsymbol{W}_{\mathrm{q}}(s)\right)=\frac{(\lambda m_1)^2\rho}{\lambda s!(1-\rho)^2}\Omega.$$

Hence for the M/M/s model we obtain

$$\mathrm{CW}^{\exp}(s)=\frac{\mathscr{E}^{\exp}\left(\boldsymbol{W}_{\mathrm{q}}(s)\right)}{P_W(\exp)}=\frac{\rho}{\lambda(1-\rho)}$$

and this implies

$$s_{\mathrm{OPT}}=\left\lceil\lambda m_1+\frac{m_1}{D}\right\rceil\approx\lambda m_1+\frac{m_1}{D},\tag{6}$$

with $\lceil x\rceil$ denoting the smallest integer bigger than or equal to x.

However, for arbitrary distributions G it is impossible to obtain an exact expression for $\mathrm{CW}(s)$ (cf. [4,19]) and so we have to use some approximation formula. Denote by $p_n(\text{app})$,

$\mathscr{E}^{\text{app}}(W_q(s))$ and $P_W(\text{app})$ the approximations of resp. p_n, $\mathscr{E}(W_q(s))$ and P_W. In the literature (cf. [2,4,19]) there are several approximations available for the mean waiting time and the delay probability. An attractive one for the mean waiting time based on theoretical assumptions and very accurate in case $c_s^2 := (m_2/m_1^2) - 1 \leq 1$ (relative error percentage between 0 and 2%) is given by (cf. [2,4,19,20]).

$$\begin{aligned}\mathscr{E}^{\text{app}}(W_q(s)) &= \rho\frac{m_2}{2m_1^2}\mathscr{E}^{\text{exp}}(W_q(s)) \\ &\quad + (1-\rho)\frac{s\gamma_1}{m_1}\mathscr{E}^{\text{exp}}(W_q(s)), \end{aligned} \tag{7}$$

with

$$\gamma_1 := \int_0^\infty (1 - G_e(z))^s \, dz$$

and

$$G_e(z) := \frac{1}{m_1}\int_0^z (1 - G(t)) \, dt.$$

Note that the above approximation is a convex combination of the light and heavy traffic results available for the M/G/s queue (cf. [1,6]) and is consistent with those. Moreover, by the same theoretical assumptions one can derive (cf. [4,19])

$$P_W(\text{app}) = P_W(\text{exp}). \tag{8}$$

Hence by (7) and (8),

$$\begin{aligned}\text{CW}^{\text{app}}(s) &= \frac{\mathscr{E}^{\text{app}}(W_q(s))}{P_W(\text{app})} \\ &= \rho\frac{m_2}{2m_1^2}\text{CW}^{\text{exp}}(s) \\ &\quad + (1-\rho)\frac{s\gamma_1}{m_1}\text{CW}^{\text{exp}}(s) \\ &= \frac{\lambda m_2}{2s(s - \lambda m_1)} + \gamma_1. \end{aligned} \tag{9}$$

In order to determine $s_{\text{OPT}}(\text{app})$ we need an upper and lower bound on γ_1. These bounds are provided in the next lemma.

Lemma 2. *For any distribution G satisfying* $G(0^+) < 1$ *we have* $\gamma_1 > m_1/(s+1)$. *Also, if G is some* NBUE (*New Better than Used in Expectation*)-*distribution* (*that is* $1 - G_e(t) \leq 1 - G(t)$ $\forall t > 0$) *it follows that* $\gamma_1 \leq m_1/s$.

Proof. Clearly

$$\begin{aligned}&-\frac{(s+1)}{m_1}\int_0^\infty (1 - G_e(t))^s(1 - G(t)) \, dt \\ &= \int_0^\infty d\big((1 - G_e(t))^{s+1}\big) = -1\end{aligned}$$

and this yields

$$\begin{aligned}&\int_0^\infty (1 - G_e(t))^s \, dt \\ &\quad > \int_0^\infty (1 - G_e(t))^s(1 - G(t)) \, dt = \frac{m_1}{s+1}.\end{aligned}$$

For the reverse inequality we note whenever G is some NBUE-distribution that

$$\begin{aligned}&\int_0^\infty (1 - G_e(t))^s \, dt \\ &\quad \leq \int_0^\infty (1 - G_e(t))^{s-1}(1 - G(t)) \, dt = \frac{m_1}{s}. \quad \square\end{aligned}$$

In the remainder we only consider service times having a distribution of the above type. (For a detailed discussion on NBUE-distributions the reader is referred to [13]). If this assumption is too strict for certain shop types we approximate γ_1 by

$$\gamma_1(\text{app}) = \left(1 - c_s^2\right)\frac{m_1}{s+1} + c_s^2\frac{m_1}{s}$$

(cf. [19]) and use a similar type of analysis for the derivation of the relation between labour volume and sales.

By lemma 2 and (9) we obtain

$$\begin{aligned}&\frac{\lambda m_2}{2s(s - \lambda m_1)} + \frac{m_1}{s+1} \\ &\quad < \text{CW}^{\text{app}}(s) \leq \frac{\lambda m_2}{2s(s - \lambda m_1)} + \frac{m_1}{s} \end{aligned} \tag{10}$$

and this yields after some calculations, having in mind that

$$s_{\text{OPT}}(\text{app}) := \inf\{s \colon \text{CW}^{\text{app}}(s) \leq D\}$$

and

$$\frac{\lambda m_2}{2(s+1)(s+1-\lambda m_1)} + \frac{m_1}{s+1}$$

is by (10) also a lower bound for $\text{CW}^{\text{app}}(s)$, that $s_{\text{OPT}}(\text{app})$ is bounded above by

$$\left\lceil \tfrac{1}{2}\left(\lambda m_1 + \frac{m_1}{D}\right) + \tfrac{1}{2}\left(\left(\lambda m_1 - \frac{m_1}{D}\right)^2 + \frac{2\lambda m_2}{D}\right)^{1/2} \right\rceil$$

and bounded below by

$$\left[\tfrac{1}{2}\left(\lambda m_1 + \frac{m_1}{D}\right) + \tfrac{1}{2}\left(\left(\lambda m_1 - \frac{m_1}{D}\right)^2 + \frac{2\lambda m_2}{D}\right)^{1/2} - 1\right].$$

Since the difference between the upper and lowerbound is 1, we take the optimal number of cashiers equal to

$$\begin{aligned} s_{\mathrm{OPT}} &\simeq s_{\mathrm{OPT}}(\mathrm{app}) \\ &\simeq \tfrac{1}{2}\left(\lambda m_1 + \frac{m_1}{D}\right) \\ &\quad + \tfrac{1}{2}\left(\left(\lambda m_1 - \frac{m_1}{D}\right)^2 + \frac{2\lambda m_2}{D}\right)^{1/2}. \end{aligned} \tag{11}$$

(Note for the M/M/1 model we have $s_{\mathrm{OPT}} = s_{\mathrm{OPT}}(\mathrm{app})$!) Hence

$$\begin{aligned} \mathscr{E}\boldsymbol{G}(s_{\mathrm{OPT}}) &\simeq \mathscr{E}\boldsymbol{G}(s_{\mathrm{OPT}}(\mathrm{app})) = s_{\mathrm{OPT}}(\mathrm{app})T \\ &\quad + \mathscr{E}\left(\max\left(\gamma \sum_{i=1}^{N_T} \boldsymbol{Q}_i \right.\right. \\ &\qquad - \mathscr{E}(\xi(\rho(s_{\mathrm{OPT}}(\mathrm{app})))) \\ &\qquad \left.\left. \times \boldsymbol{I}(s_{\mathrm{OPT}}(\mathrm{app})), 0\right)\right). \end{aligned} \tag{12}$$

Clearly by the Poisson arrival process the expected total sales are given by

$$Q = \mathscr{E}\left(\sum_{i=1}^{N_T} \boldsymbol{Q}_i\right) = \lambda q_1 T \tag{13}$$

with $q_1 := \mathscr{E}(\boldsymbol{Q}_i)$, $i \geq 1$, and hence by (11) and (13),

$$\begin{aligned} s_{\mathrm{OPT}} \simeq s(Q) &:= \tfrac{1}{2}\left(\frac{Qm_1}{q_1 T} + \frac{m_1}{D}\right) \\ &\quad + \tfrac{1}{2}\left(\left(\frac{Qm_1}{q_1 T} - \frac{m_1}{D}\right)^2 + \frac{2Qm_2}{q_1 TD}\right)^{1/2}. \end{aligned} \tag{14}$$

Also, by definition $\mathscr{E}\boldsymbol{G}(s_{\mathrm{OPT}})$ denotes the expected labour volume L and by (12) and (14) we obtain the following inequality between L and Q

$$F_2(Q) \leq L \leq F_1(Q), \tag{15}$$

with

$$\begin{aligned} F_1(Q) &:= \tfrac{1}{2}\left(\frac{m_1 Q}{q_1} + \frac{m_1 T}{D}\right) \\ &\quad + \tfrac{1}{2}\left(\left(\frac{Qm_1}{q_1} - \frac{m_1 T}{D}\right)^2 + \frac{2QTm_2}{q_1 D}\right)^{1/2} \\ &\quad + \gamma Q \end{aligned} \tag{16}$$

and

$$F_2(Q) := F_1(Q) - \mathscr{E}(\xi(\rho(s(Q))))\mathscr{E}(\boldsymbol{I}(s(Q))). \tag{17}$$

We are now able to prove our main result which is asymptotically consistent with the empirically verified non-homogeneous relation between sales and labour volume whenever D equals $m_2/2m_1$.

Theorem 3. *Suppose shopkeepers face an optimization problem as discussed in* (5). *If the service requirements of customers are given by some* NBUE-*distribution then with L denoting labour volume and Q total sales, we obtain*

$$L = Q\left(\gamma + \frac{m_1}{q_1}\right) + \frac{m_2 T}{2m_1 D} + \mathrm{o}(1), \tag{18}$$

where o(1) *denotes any function $f(Q)$with* $\lim_{Q \to \infty} f(Q) = 0$.

Proof. First we have to prove that $L - Q(\gamma + m_1/q_1) - m_2T/(2m_1D)$ is uniformly bounded on $[0, \infty)$.

By (16) it follows immediately that

$$\begin{aligned} &F_1(Q) - Q\left(\gamma + \frac{m_1}{q_1}\right) \\ &\quad = \tfrac{1}{2}\left(\left(\frac{Qm_1}{q_1} - \frac{m_1 T}{D}\right)^2 + \frac{2QTm_2}{q_1 D}\right)^{1/2} \\ &\qquad - \tfrac{1}{2}\left(\frac{Qm_1}{q_1} - \frac{m_1 T}{D}\right) \end{aligned}$$

Observe for any $y > 0$,

$$\tfrac{1}{4}y(x^2 + y)^{-1/2} \leq \tfrac{1}{2}(x^2 + y)^{1/2} - \tfrac{1}{2}x \leq \tfrac{1}{4}\frac{y}{x} \tag{19}$$

and hence substituting $x = (Qm_1/q_1) - (m_1T/D)$ and $y = 2QTm_2/q_1D$ yields

$$\begin{aligned}
&\tfrac{1}{2}QTm_2\left((Qm_1D - m_1q_1T)^2 + 2q_1DQTm_2\right)^{-1/2}\\
&\quad \le F_1(Q) - Q\left(\gamma + \frac{m_1}{q_1}\right)\\
&\quad \le \tfrac{1}{2}QTm_2(Qm_1D - q_1m_1T)^{-1}. \qquad (20)
\end{aligned}$$

The above inequalities easily imply

$$\lim_{Q\to\infty} F_1(Q) - Q\left(\gamma + \frac{m_1}{q_1}\right) - \frac{Tm_2}{2m_1D} = 0. \qquad (21)$$

Moreover, if

$$f(j) := \begin{cases} 0 & \text{for } j \ge s(Q), \\ s - j & \text{for } 0 \le j < s(Q), \end{cases}$$

we obtain by the definition of the random variable $\boldsymbol{I}(s(Q))$ that

$$\begin{aligned}
\mathscr{E}(\boldsymbol{I}(s(Q))) &= \mathscr{E}\left(\int_0^T f(X_t^*)\,\mathrm{d}t\right)\\
&= T\mathscr{E}\ (\text{number of idle servers})
\end{aligned}$$

and hence by Little's formula and the boundedness of ξ

$$\begin{aligned}
&\mathscr{E}(\xi(\rho(s(Q))))\mathscr{E}(\boldsymbol{I}(s(Q)))\\
&\quad \le \mathscr{E}(\boldsymbol{I}(s(Q)))\\
&\quad = T(s(Q) - \lambda m_1) = T\left(s(Q) - \frac{Qm_1}{q_1T}\right). \qquad (22)
\end{aligned}$$

By (21), (22) and the relations (14)–(17) the uniform boundedness is verified. Finally we have to prove that

$$\lim_{Q\to\infty} L - Q\left(\gamma + \frac{m_1}{q_1}\right) - \frac{m_2T}{2m_1D} = 0.$$

In order to verify this we note by the monotonicity of ξ that

$$\begin{aligned}
&\mathscr{E}(\xi(\rho(s(Q))))\\
&\quad = \mathscr{E}\left(\xi\left(1 - \frac{I(s(Q))}{s(Q)T}\right)\right)\\
&\quad = \mathscr{E}\left(\xi\left(1 - \frac{I(s(Q))}{s(Q)T}\right)1_{\{I(s(Q)) > \varepsilon s(Q)T\}}\right)\\
&\qquad + \mathscr{E}\left(\xi\left(1 - \frac{I(s(Q))}{s(Q)T}\right)1_{\{I(s(Q)) \le \varepsilon s(Q)T\}}\right)\\
&\quad \le P\{\boldsymbol{I}(s(Q)) > \varepsilon s(Q)T\} + \xi(1 - \varepsilon), \qquad (23)
\end{aligned}$$

for any $\varepsilon > 0$. Since by (14), $\lim_{Q\to\infty}(s(Q)q_1T/Qm_1) = 1$, we obtain by Markov's inequality and (22)

$$\begin{aligned}
&P\{\boldsymbol{I}(s(Q)) > \varepsilon s(Q)T\}\\
&\quad \le \frac{\mathscr{E}(\boldsymbol{I}(s(Q)))}{\varepsilon s(Q)T} = \frac{T\left(s(Q) - \dfrac{Qm_1}{q_1T}\right)}{\varepsilon s(Q)T}\\
&\quad = \frac{1}{\varepsilon} - \frac{1}{\varepsilon}\frac{Qm_1}{q_1Ts(Q)} \to 0 \quad (Q \to \infty)
\end{aligned}$$

and hence by (23) and the continuity of ξ

$$\lim_{Q\to\infty} \mathscr{E}(\xi(\rho(s(Q)))) = 0.$$

This yields

$$\lim_{Q\to\infty} \mathscr{E}(\xi(\rho(s(Q))))\mathscr{E}(I(s(Q))) = 0$$

and by (21) and (15)–(17) the desired result follows. □

3. Conclusion

We have proved in this paper that the empirically verified linear relationship between labour volume and sales can be explained by some optimization problem. However, it is clear that some of the assumptions of the underlying queueing model are not realistic. The main deficit is the Poisson arrival process which need to be replaced by some time-inhomogeneous Poisson process. However, in this case we do not know of any reasonable approximation for the expected waiting time provided the queue becomes stable eventually. Another deficit is the lack of micro-data to estimate the form of the service time distribution. We suspect that at least for some shop types an NBUE-distribution is a good fit. Both remarks will be subject to further research.

References

[1] Burman, D.Y., and Smith, D.R., "A light traffic theorem for multi-server queues", *Mathematics of Operations Research* 8 (1983) 15–25.

[2] Groeneveldt, H., van Hoorn, M.H., and Tijms, H.C., "Tables for M/G/c systems with phase-type service", *European Journal of Operational Research* 16 (1984) 257–269.

[3] Gross, D., and Harris, C.M., *Fundamentals of Queueing Theory*, Wiley, New York, 1985.

[4] van Hoorn, M.H., "Algorithms and approximations for queueing systems", CWI Tract No. 8, CWI, Amsterdam, 1984.

[5] Kelly, F.P., *Reversibility and Stochastic Networks*, Wiley, New York, 1979.

[6] Köllerström, J., "Heavy traffic theory for queues with several servers", *Journal of Applied Probability* 11 (1974) 544–552.

[7] N.M. Mirasol, The output of an $M/G/\infty$ queueing system is Poisson, *Operations Research* 11 (1963) 282–284.

[8] Nooteboom, B., *Retailing: Applied Analysis in the Theory of the Firm*, J.C. Gieben, Amsterdam/Uithoorn, 1980.

[9] Nooteboom, B., "Threshold costs in service industries", *Service Industries Journal* 6/1 (1987) 65–76.

[10] Nooteboom, B., "Productivity growth in the grocery trade", *Applied Economics* 15 (1983) 649-664.

[11] Nooteboom, B., "Costs, margins and competition: Causes of structural change in retailing", *International Journal of Research in Marketing* 3 (1986) 233–242.

[12] Nooteboom, B., "A new theory of retailing costs", *European Economic Review* 17 (1982) 163–186.

[13] Ross, S.M., *Stochastic Processes*, Wiley, New York, 1983.

[14] Stoyan, D., *Comparison Methods for Queues and Other Stochastic Models*, Wiley, New York, 1983.

[15] Thurik, A.R., "Labour productivity, economics of scale and opening time in large retail establishments", *Service Industries Journal* 4/1 (1984) 19–29.

[16] Thurik, A.R., "Transaction per customer in supermarkets", *International Journal of Retailing* 1/3 (1986) 33–42.

[17] Thurik, A.R., and van der Wijst, D., "Part-time labour retailing", *Journal of Retailing* 60/3 (1984) 62–80.

[18] Thurik, A.R., and Kleijweg, A., "Procyclical retail labour productivity", *Bulletin of Economic Research* 38/2 (1986) 169–175.

[19] Tijms, H.C., *Stochastic Modelling and Analysis*, Wiley, New York, 1986.

[20] Tijms, H.C., and van Hoorn, M.H., "Computational methods for single-server and multi-server queues with Markovian input and general service times" in: R.L. Disney, T.J. Ott (eds.), *Applied Probability Computer Science, the interface*, Vol. II, Birkhauser, Boston, 1982, pp. 71–102.

[21] Whitt, W., "Embedded renewal processes in the $GI/G/s$ queue", *Journal of Applied Probability* 9 (1972) 650–658.