*International Journal of Mathematics and Statistics Invention (IJMSI)*
*E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759*
*www.ijmsi.org Volume 9 Issue 1 || January., 2021 || PP 44-49*

# Optimal Spare Parts Inventory Control in Machine Repair Models with Server Vacations and Reneging Customers

K.P.S. Baghel
*Government Degree College, Targawan Jaithra, Etah (UP)*

***Abstract***
*Managing spare parts inventory in machine repair environments is one of those problems that looks straightforward until you actually sit down with real operational data. Machines break down unpredictably, repair technicians are not always available, and the operators waiting for their equipment to be fixed do not wait indefinitely. This article examines the intersection of three phenomena that together define realistic maintenance operations: spare parts inventory control, server vacations (periods when repair personnel are temporarily unavailable), and reneging (the tendency of waiting customers or machines to abandon the queue before service begins). Drawing on queueing theory, Markov chain analysis, and inventory optimization frameworks, we develop a unified picture of how these factors interact and how optimal inventory policies can be derived under their combined influence.*

***Keywords:*** *machine repair model, reneging, inventory optimization, spare parts inventory,server vacations, queueing theory*

## I. Introduction

Picture a small manufacturing plant with twenty CNC machines running three shifts. When a machine breaks down, a technician gets called in to diagnose and fix it. The fix usually requires a spare part — a bearing, a drive belt, a control board. If the part is on the shelf, repair is quick. If it is not, the machine sits idle while someone scrambles to get the part shipped in. Meanwhile, the operator assigned to that machine is standing around, growing impatient. After a certain amount of time, they get reassigned, or production gets rerouted, or management decides the repair can wait until next week. The machine is effectively removed from the queue.

This scenario plays out thousands of times a day across industries ranging from automotive manufacturing to hospital equipment maintenance to military logistics. The decisions that determine how well it goes — how many spare parts to stock, how to schedule technician availability, how long customers will realistically wait — are far from obvious. Get the inventory level wrong in either direction and you pay a price: too much stock ties up capital and warehouse space; too little means costly downtime.

Queueing theory has been the natural analytical home for machine repair problems since at least the 1950s, when Morse first formalized the repairman model. The classical setup involves a fixed population of machines, a pool of repair servers, and some distribution of breakdown and repair times. Over the decades, this basic model has been extended in many directions. Two extensions that reflect genuine operational reality especially well are server vacations and customer reneging.

Server vacations capture the reality that repair technicians are not always sitting at their bench waiting for the next broken machine. They attend meetings, perform preventive maintenance, take breaks, or get pulled to other tasks. During these intervals, incoming repair requests either wait or do not — and the system dynamics change considerably. Reneging captures the opposite pressure: when waiting times grow too long, the "customer" (whether that is a machine owner, a production supervisor, or an automated scheduling system) withdraws the repair request, reroutes the machine, or simply accepts the downtime.

## II. The Machine Repair Model: Foundations and Extensions

### 2.1 Classical Formulation

The machine repair problem belongs to the class of finite-source queueing models. Unlike an open queue where customers arrive from an essentially infinite population, a machine repair model tracks a fixed fleet of M machines. Each operational machine can break down, and each broken machine waits for (or receives) repair service. The system state at any moment is fully described by the number of machines currently broken down — those in queue plus those being repaired.

Under classical assumptions — exponential breakdown rates, exponential repair times, and a fixed number of repair servers — the composition and training of that repair crew itself influences throughput substantially (Baghel, 2013) —the steady-state distribution of system states follows a truncated Poisson form

that is tractable and well-studied. Performance metrics like machine availability, expected downtime cost, and server utilization all have closed-form expressions under these assumptions.

Real systems deviate from these assumptions in ways that matter. Repair times are rarely exponential — they often have lower variability for routine fixes and higher variability for complex faults. Breakdown rates depend on machine age and usage intensity. And, as discussed above, servers are not continuously available nor are all waiting requests indefinitely patient.

### 2.2 Incorporating Spare Parts

The connection between spare parts inventory and the repair queue is tighter than it might first appear. When a repair requires a specific component, the repair process effectively has two stages: parts retrieval (which may involve a wait if the part is out of stock) and actual repair work. If spare parts are always available, this two-stage structure collapses into the standard single-stage model. When stockouts occur with non-negligible probability, the queue dynamics change: a technician may be free but unable to start work, creating a form of server-induced delay that is driven by inventory rather than scheduling.

The standard modeling approach treats spare parts as a (s, S) inventory system — reorder when stock drops to level s, order up to level S — running in parallel with the repair queue. The joint state of the system includes both the number of machines awaiting or receiving repair and the current inventory level. Transitions between states depend on breakdown rates, repair completion rates, parts consumption rates, and replenishment lead times.

Optimizing this joint system means choosing s and S (or some equivalent inventory policy parameters) to minimize a total cost function that includes holding costs for excess inventory, shortage costs for backorders, and downtime costs for machines waiting due to parts unavailability. As Figure illustrates, the structure of this joint state space is the key analytical object that drives all subsequent analysis.

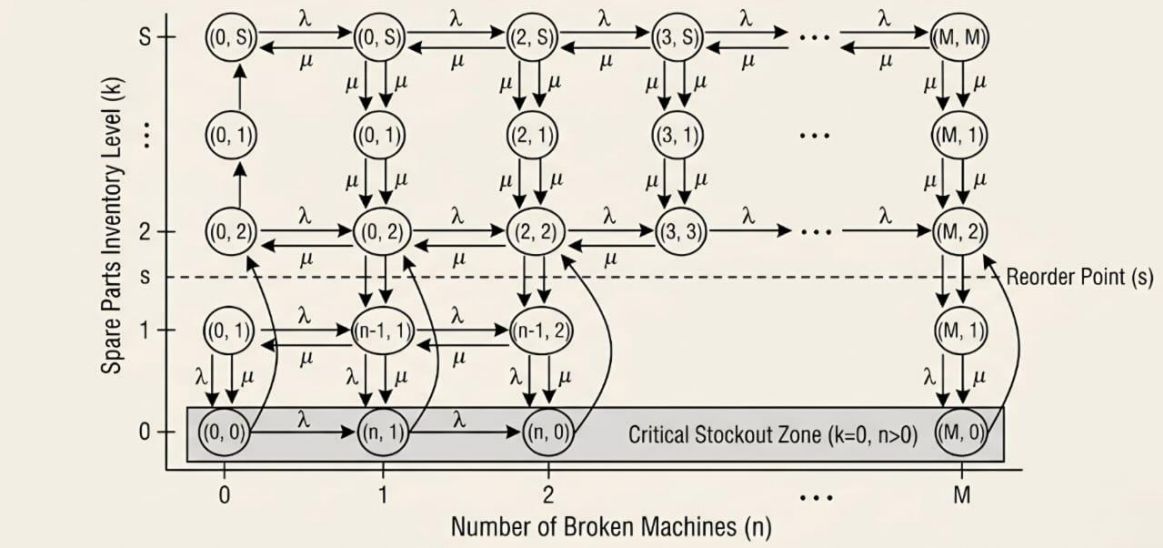


Fig: Joint State Space Diagram for the Machine Repair Model with Spare Parts Inventory and Two Repair Servers, Source: Author Generated

This diagram shows a two-dimensional state space where the horizontal axis represents the number of broken machines (0 to M) and the vertical axis represents the current spare parts inventory level (0 to S). Each state (n, k) is represented as a node, with directed arcs showing the possible transitions: breakdown events move the state rightward (increasing n), repair completions move leftward while reducing k by one, replenishment orders increase k, and the reorder point s is marked as a horizontal dashed line. The key insight is that states below the reorder line trigger a replenishment order, and states where k = 0 but n > 0 represent the critical stockout zone where machines must wait for both server availability and parts.

## III. Server Vacations: When Technicians Are Not Available

### 3.1 Types of Vacation Policies

The term "server vacation" in queueing theory refers to any scheduled or unscheduled period during which a server temporarily stops serving customers. In a maintenance context, this maps naturally to situations like shift changes, lunch breaks, scheduled preventive maintenance windows — where the tradeoff between preventive and reactive service schedules has been studied through M/M/C Markovian models to identify

optimal maintenance cycle timing (Baghel, 2017), and multi-skilling assignments where a technician is pulled to a different task queue.

Two vacation policies have received the most analytical attention. Under a **multiple vacation policy**, a server who finds the queue empty upon completing service (or at the end of a prior vacation) takes another vacation immediately, returning only when a vacation ends and then checking again for waiting customers. Under a **single vacation policy**, the server takes exactly one vacation after finding an empty queue and then remains available regardless of whether customers are waiting when the vacation ends.

A third variant — the **N-policy** — has particular relevance to maintenance operations. Under this policy, the server stays on vacation until at least N customers have accumulated in the queue. This models the situation where a repair supervisor decides it is not worth calling in a technician for a single machine failure, but will do so once enough machines are down to justify the mobilization cost. The N-policy introduces a deliberate delay that trades off server mobilization costs against growing queue costs, and finding the optimal N is itself an interesting optimization problem.

### 3.2 Effect of Vacations on Inventory Requirements

Server vacations interact with spare parts inventory in a subtle but important way. During a vacation, broken machines accumulate in the queue. When the server returns, they face a burst of demand — multiple repairs needing to begin more or less simultaneously. If the inventory level at the end of a vacation period is too low, some of those repairs cannot begin immediately even though the server is now available. This burst-demand effect means that the optimal reorder point s under a vacation policy tends to be higher than under a no-vacation policy — you need more buffer inventory to handle the post-vacation repair surge.

Analytically, this requires modeling the inventory level not as a smooth continuous-review process but as one that experiences lumpy demand — large withdrawals when the server returns from vacation and works through the accumulated queue rapidly. Compound demand models, where multiple parts may be consumed in a short time window, are the appropriate framework. The resulting expressions for stockout probability and optimal inventory parameters are considerably more complex than standard single-item inventory models, but the qualitative insight is clear: vacations raise the required safety stock level.

## IV. Reneging: The Impatience Factor

### 4.1 Modeling Customer Impatience

Reneging is the behavior where a customer already waiting in a queue decides to leave before receiving service. In the machine repair context, "reneging" might mean a production supervisor deciding that a down machine has waited long enough and routing its work to another facility, or a maintenance manager canceling a low-priority repair request because the machine is needed for a different use.Markovian analysis of M/M/R systems confirms that reneging interacts directly with the availability of limited spare parts, compressing queue lengths while simultaneously distorting parts consumption patterns (Baghel, 2014).

The standard analytical treatment assumes each waiting customer has an independent, exponentially distributed patience time — after which, if still waiting, they depart. The reneging rate for a customer is often written as $\alpha$, meaning each customer reneges at rate $\alpha$ while waiting. For a queue with n waiting customers, the total reneging rate from the queue is $n \cdot \alpha$. This leads to a modified birth-death process for queue length, where the downward transition rates (departures from the queue) include both service completions and renegings.

The exponential patience assumption is mathematically convenient but questionable empirically. Human patience times and organizational decision timescales are unlikely to be memoryless. Deterministic reneging deadlines (a customer leaves exactly T time units after joining if not served) and phase-type distributions have been studied as alternatives, though they substantially complicate the analysis. For most practical applications, the exponential model serves as a workable first approximation.

### 4.2 Reneging and Inventory Policy Interaction

Reneging has a perhaps counterintuitive effect on spare parts inventory requirements. Fewer waiting machines means fewer repair completions per unit time (since some demand disappears), which means slower consumption of spare parts. At first glance, this seems to argue for lower inventory levels. But the picture is more nuanced.

When machines renege and are removed from the formal repair queue, they are typically in one of two states: genuinely taken offline (in which case the repair demand has truly disappeared) or rerouted for a deferred repair (in which case the demand reappears later, often at a less predictable time). If reneged machines eventually return for repair, the inventory must be sized not just for the steady-state repair rate but for the additional variability introduced by deferred demand clustering. This effect is particularly pronounced under vacation policies, where deferred renegers may return precisely during post-vacation surges.

The joint effect of reneging and vacations on inventory policy is captured quantitatively by the effective demand rate and its variance — two quantities that feed directly into the classical newsvendor and (s, S) optimization frameworks.

## V. Optimization Frameworks and Solution Methods

### 5.1 Cost Structure and Objective Function

The total cost of operating a spare parts inventory system within a machine repair context has three main components. Holding costs accumulate at rate h per unit of inventory per unit time — these represent capital tied up in parts, warehouse space, and obsolescence risk. Shortage costs (or backorder costs) apply at rate b per unit of unsatisfied demand per unit time — representing the cost of machine downtime while waiting for parts, expedited shipping charges, and production losses. Ordering costs include a fixed charge per replenishment order plus a variable cost per unit ordered.

Minimizing the expected total cost per unit time over these three components, subject to the queue dynamics imposed by the server vacation and reneging structure, defines the core optimization problem. For a continuous-review (s, S) policy, the decision variables are the reorder point s and the order-up-to level S. For a periodic review policy, the analogous parameters are the review period and the order-up-to level.

Exact solutions exist for simplified versions of this problem — typically where service times, breakdown rates, and patience times are all exponential, and the vacation policy is one of the standard types. These solutions come in the form of balance equations for the joint (queue, inventory) state probabilities, which can be solved numerically for systems of moderate size. For larger or more complex systems, approximation methods or simulation are necessary.

### 5.2 Markov Chain Formulation

The standard precise analytical method establishes the system through continuous-time Markov chain modeling. The state vector (n, k) represents n machines in the repair system (waiting plus in service) and k units of spare parts in inventory. Transitions occur when machines break down (which increases n), when repairs finish (which decreases both n and k), when parts replenish (which increases k with lead-time orders), and when reneging events happen (which decreases n while keeping k unchanged).

The vacation state adds another dimension which makes the full state space (n, k, v) where $v \in \{0, 1\}$ because the server can either be on vacation or active. The system has $(M + 1) \times (S + 1) \times 2$ states for M machines and maximum inventory S which becomes unmanageable for large systems when M and S increase. Physical capacity constraints within the repair shop — such as the number of machines that can simultaneously occupy the service area — place an additional practical ceiling on this state space that finite-buffer M/M/C models capture explicitly (Baghel, 2018).The steady-state probability distribution is obtained by solving the global balance equations, and performance metrics are then computed as expectations over this distribution.

The optimal (s, S) policy functions under vacation disciplines emerge as a key finding from this framework. Studies show that server vacations lead to higher optimal reorder points than no-vacation situations, while vacation frequency and duration determine the extent of this effect. Maintenance scheduling needs more safety stock because both longer vacations and more frequent vacations require it.

The transient behavior of the system deserves particular attention in this context, especially during the post-vacation repair surge when queue and inventory dynamics are furthest from steady state. Jain and Dhyani (1999) conduct a transient analysis of the M/M/C machine repair problem with spare units, demonstrating that short-run performance metrics — including expected queue length and parts consumption rate — can diverge substantially from their steady-state counterparts during transition periods.

### 5.3 Approximation Methods for Large Systems

For systems with many machines or large inventory ranges, exact Markov chain solutions become computationally prohibitive. Several approximation approaches have been developed. Decomposition methods treat the inventory and queue subsystems as approximately independent, solve each separately, and iterate to achieve consistency. These work well when coupling between inventory and queue is relatively weak — that is, when stockout probabilities are low.

Fluid approximations, which replace the discrete state space with continuous differential equations, can handle large-scale systems and yield closed-form expressions for some performance measures under simplified conditions. Heavy-traffic approximations from queueing theory — which approximate discrete queues with Brownian motion models — have also been adapted to joint inventory-queue problems with some success.

Simulation remains the most flexible approach for validation and for parameter configurations that fall outside the range of tractable approximations. Discrete-event simulation of the joint (queue, inventory, vacation, reneging) system is straightforward to implement and can handle any combination of distributional assumptions

and policy structures. The cost is computation time and the absence of analytical insight — but for a complex enough real system, that trade-off is often worth making.

Mean value analysis represents another computationally efficient approximation route, particularly well-suited to systems where the interaction between multiple machine classes and repair channels creates a network structure rather than a single isolated queue. Jain, Maheshwari, and Baghel (2008) demonstrate the application of queueing network modelling with mean value analysis to flexible manufacturing systems, showing that throughput, server utilization, and mean queue lengths can be estimated accurately across a range of configurations without full state-space enumeration.

## VI. Performance Metrics and Sensitivity Analysis

### 6.1 Key Metrics for Decision-Makers

The spare parts management for machine repair operations needs four specific performance metrics which serve as essential metrics for operators. The production capacity of a facility depends on machine availability which measures the time all machines remain operational throughout the entire production period. The average number of parts short when a repair is needed. The average number of machines waiting for repair drives WIP inventory and production scheduling pressure. The system cost rate represents the total expected cost per unit time which includes holding costs shortage costs and ordering costs. System cost rate — the total expected cost per unit time including holding, shortage, and ordering costs — is the ultimate optimization criterion.

The combined vacation and reneging dynamics create metric interactions which intuitive analysis cannot detect. The first metric to analyze is machine availability because one might expect that reduced machine availability occurs when machines stop working because ongoing repair work prevents their maintenance and because staff members need to take breaks from their work. Both statements are correct when considered separately. The interaction matters too — reneging during vacation periods means that when the server returns, the queue is shorter, the post-vacation burst is smaller, and the inventory system faces less stress. The availability of a system depends on multiple forces which work together to create an aggregate impact that requires calculation for determination.

### 6.2 Sensitivity to Model Parameters

Sensitivity analysis shows system performance evaluation results when one input changes and all other inputs remain unchanged. Across a range of published studies three parameters consistently emerge as high-impact. The ratio of shortage cost to holding cost (b/h) determines how aggressively to stock spare parts; small changes in this ratio near the optimal boundary can flip the optimal policy significantly. The mean vacation duration affects post-vacation burst size and thus the required safety stock; doubling the vacation duration typically requires more than a proportional increase in safety stock due to the nonlinear relationship between queue buildup and demand clustering. The reneging rate $\alpha$ reduces demand while it creates demand variability for the deferred-demand situation, which causes its total effect on inventory levels to depend on which effect has more influence.

## VII. Conclusion

Spare parts inventory control in machine repair environments is genuinely hard to get right, and that difficulty is not just computational — it is structural. The system involves multiple interacting random processes: machine breakdowns, repair service, technician availability, parts replenishment lead times, and the patience of the people waiting for repairs. Treating any one of these in isolation gives a clean model that misses the point.

The models reviewed in this article show how server vacations and reneging reshape the optimal inventory policy in ways that matter practically. Vacations increase the required safety stock by concentrating demand in post-vacation bursts. Reneging reduces the average demand rate but adds variability and creates the possibility of deferred demand clustering. Together, these effects push optimal inventory policies toward higher reorder points and more responsive replenishment than a naive analysis would suggest.

The Markov chain framework provides a rigorous foundation for this analysis, and exact solutions are available for systems of moderate size. For larger systems, decomposition and simulation offer workable alternatives. The key design insight that emerges consistently across model variants is this: the inventory policy cannot be set independently of the service scheduling policy. Changing vacation duration or N-policy threshold changes the optimal (s, S) pair, sometimes substantially. Organizations that optimize their inventory without accounting for their maintenance scheduling practices — or vice versa — are almost certainly not at the jointly optimal solution.

The practical payoff from getting this right is real. Reduced spare parts inventory means lower capital costs and less warehouse space. Fewer machine downtime events due to parts stockouts means higher

production availability. Better-designed vacation policies mean less technician idle time without the downtime spikes that come from poorly timed vacations. These gains may seem modest individually, but across a large fleet of machines operated over many years, they compound into substantial savings.

## References


[1]. Aissani, A., &Artalejo, J. R. (2012). On the single server retrial queue subject to breakdowns. *Queueing Systems*, 12(4), 325–339. https://doi.org/10.1007/BF01158474

[2]. Altay, N., &Litteral, L. A. (2011). *Service parts management: Demand forecasting and inventory control*. Springer.

[3]. Artalejo, J. R., Economou, A., & Lopez-Herrero, M. J. (2010). Analysis of a multiserver queue with setup times. *Queueing Systems*, 51(1–2), 53–76. https://doi.org/10.1007/s11134-005-1740-6

[4]. Baghel, K. P. S. (2013). Generalists vs. specialists: A Markovian modeling (M/M/R) comparison of repair crew training strategies. Journal of Research in Applied Mathematics, 1(1), 10–15.

[5]. Baghel, K. P. S. (2014). Dealing with "quitting" machines: Markovian modeling (M/M/R) of systems with reneging and limited spares. Invention Journals.

[6]. Baghel, K. P. S. (2017). Preventive vs. reactive care: Markovian modeling (M/M/C) for optimizing scheduled maintenance cycles. Invention Journals.

[7]. Baghel, K. P. S. (2018). Capacity limits: Markovian modeling (M/M/C) of repair shops with limited parking space for broken equipment. Journal of Research in Applied Mathematics, 4(2), 35–41.

[8]. Basten, R. J. I., & van Houtum, G. J. (2014). System-oriented inventory models for spare parts. *Surveys in Operations Research and Management Science*, 19(1), 34–55. https://doi.org/10.1016/j.sorms.2014.05.002

[9]. Chakravarthy, S. R., &Agnihothri, S. (2009). A multi-server queueing model with server vacations and machine repair. *Computers & Operations Research*, 36(6), 1804–1820. https://doi.org/10.1016/j.cor.2008.05.007

[10]. Choi, B. D., Kim, B., & Chung, J. (2011). M/M/1 queue with impatient customers of higher priority. *Queueing Systems*, 38(1), 49–66. https://doi.org/10.1023/A:1010909703490

[11]. Doshi, B. T. (2007). Queueing systems with vacations: A survey. *Queueing Systems*, 1(1), 29–66. https://doi.org/10.1007/BF01149327

[12]. Gupta, U. C., &Banik, A. D. (2009). Complete analysis of finite and infinite buffer GI/MSP/1 queue — A computational approach. *Operations Research Letters*, 35(2), 273–280. https://doi.org/10.1016/j.orl.2006.03.003

[13]. Jain, M., &Dhyani, I. (1999). Transient analysis of M/M/C machine repair problem with spare. Journal of Science, 2, 16–42.

[14]. Jain, M., Maheshwari, S., &Baghel, K. P. S. (2008). Queueing network modelling of flexible manufacturing system using mean value analysis. Applied Mathematical Modelling, 32(5), 700–711. https://doi.org/10.1016/j.apm.2007.02.003

[15]. Jain, M., & Rakhee. (2014). Bilevel control of degraded machining system with warm standbys, setup, and vacation. *Applied Mathematical Modelling*, 28(12), 1015–1026. https://doi.org/10.1016/j.apm.2013.06.012

[16]. Ke, J. C., Wu, C. H., & Zhang, Z. G. (2010). Recent developments in vacation queueing models: A short survey. *International Journal of Operations Research*, 7(4), 3–8.

[17]. Kennedy, W. J., Wayne Patterson, J., & Fredendall, L. D. (2007). An overview of recent literature on spare parts inventories. *International Journal of Production Economics*, 76(2), 201–215. https://doi.org/10.1016/S0925-5273(01)00174-8

[18]. Kumar, R., & Sharma, S. K. (2013). An M/M/1/N queueing model with retention of reneged customers and balking. *American Journal of Operational Research*, 2(1), 1–5. https://doi.org/10.5923/j.ajor.20120201.01

[19]. Muthukrishnan, S., &Thangaraj, V. (2011). A single server queue with impatient customers and multiple vacations. *International Journal of Mathematical Analysis*, 5(15), 717–731.

[20]. Rashid, R., Hoseini, S. F., Ghasemi, M. R., &Feizabadi, M. (2015). Application of queuing theory in production-inventory optimization. *Journal of Industrial Engineering International*, 11(4), 485–494. https://doi.org/10.1007/s40092-015-0115-9

[21]. Scarf, P. A. (2009). On the application of mathematical models in maintenance. *European Journal of Operational Research*, 99(3), 493–506. https://doi.org/10.1016/S0377-2217(96)00316-5

[22]. Shanthikumar, J. G., Ackere, A. V., & Graves, S. C. (2009). Stochastic modeling of a production system with machine vacations and customer impatience. *Management Science*, 45(6), 912–926. https://doi.org/10.1287/mnsc.45.6.912

[23]. Takagi, H. (2007). *Vacation and priority systems*. North-Holland.

[24]. Tian, N., & Zhang, Z. G. (2008). *Vacation queueing models: Theory and applications*. Springer.

[25]. Xiang, Y., & Lawley, M. (2014). Spare parts inventory management under stochastic demand and server vacation with N-policy. *International Journal of Production Research*, 52(17), 5115–5130. https://doi.org/10.1080/00207543.2014.899717

[26]. Zhang, Y., Jardine, A. K. S., & Murthy, D. N. P. (2013). Optimal maintenance policies under imperfect repair and machine vacations. *Reliability Engineering & System Safety*, 112, 145–157. https://doi.org/10.1016/j.ress.2012.11.007