Экономически рациональный уровень затрат на наблюдаемость в ИТ-компаниях: предельный анализ и внутренняя цена информации

Соколов Дмитрий Михайлович
аспирант, базовая кафедра цифровой экономики Института развития информационного общества, Высшая школа кибертехнологий, математики и статистики, Российский экономический университет имени Г. В. Плеханова, sokdep@yandex.ru

Актуальность исследования обусловлена тем, что выбор уровня наблюдаемости относится к задачам распределения ограниченных ресурсов фирмы между затратами на предотвращение потерь и ожидаемым ущербом от операционного риска. Затраты несёт подразделение эксплуатации, а выгоду получают бизнес-подразделения, относящиеся к другим центрам ответственности, поэтому уровень выбирается эвристически либо по остаточному принципу и систематически занижается. Целью исследования является разработка экономико-математической модели оптимизации решения о затратах на наблюдаемость, связывающей конфигурацию сбора телеметрии с денежными потерями от инцидентов, и определение экономически рационального уровня наблюдаемости и условий его реализуемости внутри фирмы. Задача сформулирована как минимизация суммы ожидаемых потерь от инцидентов и затрат на наблюдаемость по показателю покрытия. Применены микроэкономическое моделирование спроса, предельный анализ, сравнительная статика, теневые цены, трансфертное ценообразование, спецификация эконометрической идентификации, численная калибровка и метод Монте-Карло. Условие возрастания удельной стоимости минуты простоя сведено к неравенству, проверяемому по телеметрии и данным о клиентах. Раннее обнаружение образует самостоятельный канал выгоды с собственной теневой ценой, не вытекающий из выпуклости потерь. Эластичность задержек по расходам не зависит от уровня цен и при неустранимом минимуме немонотонна по покрытию; выведено правило проверки бюджета по измеримым величинам. Внутренняя цена информации устраняет недоинвестирование. Иллюстративная калибровка, не являющаяся оценкой на производственных данных, даёт рациональное покрытие 0,589, сокращение среднего времени устранения (включая обнаружение) с 60,0 до 22,2 минуты и экономию 76,1 % по сравнению с нулевым покрытием, а на канал обнаружения приходится 32,8 % предельной выгоды. Следующим шагом станет эмпирическое оценивание модели.

Ключевые слова: экономико-математическая модель, оптимизационная модель, оптимизация принятия решений, предельный анализ, затраты на наблюдаемость, стоимость простоя, операционный риск, эластичность, теневая цена, трансфертное ценообразование, эконометрическая идентификация, микросервисные архитектуры.

Введение

Под наблюдаемостью понимается способность определять состояние информационной системы по телеметрии: метрикам, журналам событий и трассировкам запросов. Выбор её уровня — частный случай распределения ограниченных ресурсов фирмы между затратами на предотвращение потерь и ожидаемым ущербом. Затраты несёт подразделение эксплуатации, а выгода реализуется в бизнес-подразделениях. В экономике информационной безопасности такое расхождение стимулов известно как источник недоинвестирования (ср. [1]), и без внутренней цены информации (о трансфертных ценах см. [2]) бюджет назначается по остаточному принципу. Время реакции на инцидент, цена которого растёт с усложнением микросервисных архитектур, измеряют средним временем обнаружения (англ. mean time to detect, MTTD) и средним временем устранения (англ. mean time to resolve, MTTR); второй показатель включает время обнаружения и далее называется временем восстановления. Совокупные издержки C_{total} складываются из потерь от инцидентов и затрат на наблюдаемость C_{obs} и минимизируются по конфигурациям сбора телеметрии *cfg*:

$$C_{\text{total}}(cfg) = C_{\text{loss}}(\text{MTTD}(cfg), \text{MTTR}(cfg)) + C_{\text{obs}}(cfg) \to \min_{cfg}. \quad (1)$$

Постановка (1) служит лишь рамкой анализа, и ни существование оптимума, ни пересчёт сокращения MTTR в деньги по постоянной удельной стоимости минуты простоя новизны не составляют.

Целью исследования является разработка экономико-математической модели оптимизации решения о затратах на наблюдаемость и определение экономически рационального уровня наблюдаемости. Объектом исследования являются процессы управления инцидентами в ИТ-компаниях с микросервисной архитектурой, рассматриваемых как экономические системы, предметом — соотношение предельных затрат на телеметрию и предельного сокращения ожидаемого ущерба, эластичность задержек и распределение выгоды между центрами ответственности. Применены микроэкономическое моделирование спроса, предельный анализ, сравнительная статика, теневые цены, трансфертное ценообразование, спецификация эконометрической идентификации, численная калибровка и метод Монте-Карло. Работа соответствует пп. 3, 4, 8 и 14 паспорта научной специальности 5.2.2.

Научная новизна состоит в следующем: 1) возрастание удельной стоимости минуты простоя выведено из поведения спроса и сведено к проверяемому условию; 2) раннее обнаружение образует самостоятельный канал выгоды с собственной теневой ценой; 3) эластичность задержек по расходам не зависит от уровня цен и при неустранимом минимуме немонотонна по покрытию; 4) выведена внутренняя (пигувианская) цена информации, устраняющая клин недоинвестирования. Числовые значения получены в иллюстративной калибровке в условных единицах (у.е.); эконометрические спецификации задаются, но не оцениваются. Статья развивает работу автора [3], где этот компромисс намечен качественно.

Современное состояние исследований

Ближайший аналог — модель рационального уровня расходов на информационную безопасность при убывающей отдаче [4]. В отечественной литературе к теме близки экспертная оценка ущерба по вербальной шкале без денежного выражения при отказе информационной системы [5], методики оценки эффективности и окупаемости таких систем [6] и многоуровневая декомпозиция экономических информационных систем [7]. С конфигурацией сбора телеметрии эти подходы не связаны. Из теории оптимального профилактического обслуживания [8] заимствован тип условия первого порядка, хотя выбирается здесь информационная конфигурация при непостоянной удельной стоимости минуты простоя. Разрыв между носителем затрат и получателем выгоды трактуется в духе многоуровневых экономических информационных систем [9]. Обзор трассировки [10] и методы локализации первопричины [11] денежного результата не дают. Неисследованными остаются кривизна ущерба, выведенная из поведения спроса, различие ценности минут обнаружения и устранения, эластичность задержек как идентифицируемый параметр и механизм децентрализованной реализации рационального уровня.

Потери от инцидента: микромодель и выпуклость

Инцидент рассматривается как реализация операционного риска. Время восстановления T (в среднем MTTR) складывается из времени до обнаружения Δ_f (в среднем MTTD) и времени диагностики и устранения Δ_d и ограничено снизу неустранимым минимумом. При пуассоновском потоке инцидентов (в среднем λ в год) со случайной тяжестью в (1) входят ожидаемые годовые потери $\lambda\mathbb{E}[C_{\text{loss}}]$, где C_{loss} — потери от одного инцидента. Параметры сбора (доля сервисов с телеметрией, частота выборки метрик, доля трассируемых запросов, срок хранения журналов) сводятся в показатель покрытия $s \in [0, 1]$,

заменяющий конфигурацию в (1). Затраты $C_{obs}(s)$ растут с ускорением, поскольку вместе с объёмом хранения растёт поток сигналов, требующих внимания дежурной смены.

Микромодель недополученного дохода. Пусть t — минуты с начала деградации, $\lambda_{tx}(t)$ — поток транзакций, m — маржа на транзакцию; штрафы по соглашению об уровне обслуживания (англ. service level agreement, SLA) учитываются отдельно. Получив отказ, клиент повторяет попытку, откладывает транзакцию (и возвращается к ней с вероятностью δ) или уходит (ср. модели с нетерпеливыми клиентами [12]). Доля ушедших $\alpha(t)$ не убывает и растёт тем быстрее, чем быстрее истощается терпение (параметр ψ); при уходе утрачивается и часть пожизненной ценности клиента (надбавка γ). Повторные попытки увеличивают нагрузку, и без вмешательства доля успешно обслуживаемых запросов $q(t)$ снижается, если спрос не падает ещё быстрее; меры после обнаружения замедляют деградацию. Удельная стоимость минуты простоя равна

$$r(t;\Delta_f) = m\,\lambda_{tx}(t)\big(1 - q(t;\Delta_f)\big)\Pi(t), \qquad \Pi(t) = \alpha(t)(1+\gamma) + (1-\alpha(t))(1-\delta), \tag{2}$$

где множитель потерь на отклонённую транзакцию $\Pi(t)$ не убывает, так как уход дороже откладывания.

Темп изменения удельной стоимости раскладывается на три измеримых слагаемых: темп сезонного спада спроса σ (со знаком минус, по счётчикам запросов), темп углубления деградации g_q (по доле успешных ответов) и темп ухода клиентов g_Π (по данным о клиентах). Условие доминирования деградации над сезонностью (D) составляет вторую часть (3):

$$\frac{r'(t;\Delta_f)}{r(t;\Delta_f)} = -\sigma(t) + g_q(t) + g_\Pi(t); \qquad \text{(D): } \sigma(t) < g_q(t) + g_\Pi(t), \quad t \in (0,T). \tag{3}$$

При (D) удельная стоимость возрастает, а накопленные потери дохода (без ступенчатых штрафов по SLA) строго выпуклы по длительности инцидента; поскольку удельная стоимость положительна, верно и обратное. Выпуклость в модель не закладывается и проверяется как гипотеза. При неубывающем спросе (D) выполняется, если хотя бы один из темпов положителен; нарушить его может падающий спрос, когда деградация и уход клиентов затухают, то есть в конце длительного инцидента или после раннего обнаружения.

Раннее обнаружение как самостоятельный канал выгоды

Минутой устранения назовём дополнительную минуту этапа Δ_d при неизменном моменте обнаружения. Штрафы p_k по SLA за превышение порогов длительности T_k одинаково удорожают её и минуту обнаружения, и выпуклость сама по себе приоритета раннему обнаружению не даёт. Он возникает потому, что после обнаружения деградация останавливается или замедляется. Пусть q_{free} — доля успешно обслуживаемых запросов без вмешательства, θ — доля темпа деградации, сохраняющаяся после обнаружения (θ = 0 при её остановке). Если деградация углубляется, меры после обнаружения её сдерживают (θ < 1), а обнаружение происходит до восстановления, то минута обнаружения дороже минуты устранения на величину

$$\frac{\partial C_{loss}}{\partial \Delta_f} - \frac{\partial C_{loss}}{\partial \Delta_d} = A = (1-\theta)\,m\int_{\Delta_f}^{T}\lambda_{tx}(t)\Pi(t)\big(-q'_{free}(\Delta_f + \theta(t-\Delta_f))\big)\,dt > 0. \tag{4}$$

Здесь A — доход, подверженный риску до восстановления и взвешенный по скорости углубления деградации. Сокращение длительности за счёт раннего обнаружения уже учтено в канале устранения (MTTR включает MTTD), поэтому λA — годовая теневая цена минуты обнаружения сверх цены минуты устранения, а их сумма — полная ценность минуты обнаружения. Надбавка растёт с длительностью устранения, так что вложения в два канала отчасти взаимозаменяемы; при θ = 1 канал обнаружения исчезает.

Предельный анализ, сравнительная статика и идентификация

Подстановка ожидаемых потерь в (1) даёт задачу выбора покрытия; внутренний оптимум s^* удовлетворяет условию первого порядка

$$\mathrm{MB}(s) \equiv -\lambda\frac{\partial \mathbb{E}[C_{loss}]}{\partial \mathrm{MTTR}}\,\mathrm{MTTR}'(s) \;\underbrace{-\lambda A\,\mathrm{MTTD}'(s)}_{\mathrm{MB}_f(s)} = C'_{obs}(s) \equiv \mathrm{MC}(s), \tag{5}$$

где слева — предельная выгода MB, справа — предельные затраты MC. Канал устранения учитывает сокращение MTTR, включая достигнутое ранним обнаружением, а канал обнаружения MB_f — надбавку λA из (4), усреднённую по тяжести инцидентов. Условие второго порядка проверяется численно.

Примем $\mathrm{MTTR} = \mathrm{MTTR}_{floor} + a\,e^{-bs}$ (для MTTD форма та же) и $C_{obs} = k\,s^p$, где k — масштаб цен. Эластичность времени восстановления по расходам E однородна нулевой степени по ценам (от k не зависит) и характеризует технологию. По покрытию она немонотонна, поскольку при малом покрытии 1 % бюджета мало добавляет к покрытию, а вблизи неустранимого минимума расходам поддаётся лишь малая часть задержки. В калибровке модуль E максимален (0,521) при покрытии около 0,88, а в рациональной точке $E = -0{,}461$, поэтому эластичность приводят вместе с покрытием, при котором она оценена. Из (5) вытекает правило проверки бюджета через эластичность ε_L ожидаемых потерь по годовому бюджету наблюдаемости S:

$$|\varepsilon_L| = \left|\frac{\mathrm{d}\ln(\lambda\mathbb{E}[C_{loss}])}{\mathrm{d}\ln s}\right| = \frac{C_{obs}(s^*)}{\lambda\,\mathbb{E}[C_{loss}](s^*)}. \tag{6}$$

В рациональной точке процент сокращения ущерба на процент прироста бюджета равен отношению бюджета к ожидаемому ущербу (в калибровке 1,23), что выявляет недо- или переинвестирование без знания всей модели. Рациональный бюджет равен 13,2 % ожидаемого ущерба при нулевом покрытии, что ниже предела Гордона и Лёба $1/e \approx 36{,}8$ % [4].

Предсказания знаков откликов. При выполненном условии второго порядка рациональное покрытие сдвигается вслед за предельной выгодой (теорема о неявной функции) и растёт с частотой инцидентов и маржой. Только микромодель (через множитель потерь в (2)), но не модели [4, 8], предсказывает рост покрытия с истощением терпения и утратой пожизненной ценности клиента и его снижение с возвратом отложенного спроса. С ростом θ доля канала обнаружения падает. Рост штрафов по SLA повышает покрытие, а влияние сокращения порогов неоднозначно, поэтому знак отклика на ужесточение SLA рассчитан и проверен методом Монте-Карло.

Стратегия идентификации. Отдача расходов убывает с покрытием, поэтому зависимость задержек от покрытия оценивается нелинейно (коэффициент панельной регрессии с фиксированными эффектами дал бы лишь среднюю отдачу). Вариацию расходов, не вызванную аварийностью, выделяет инструмент типа Бартика (внешние сдвиги расходов, взвешенные по исходным долям сервисов) [13]. Рост расходов выявляет прежде незамеченные короткие инциденты и завышает оценку отдачи (смещение отбора по обнаружению), поэтому перечень инцидентов берётся из не зависящих от телеметрии источников, а этот отбор учитывается моделью Хекмана [14].

Клин недоинвестирования и внутренняя цена информации

Задача (1) соответствует оптимуму первого ранга. Если центр затрат (эксплуатация) учитывает лишь долю κ < 1 предельной выгоды, неся все предельные затраты, его покрытие ниже s^* и тем ниже, чем меньше κ. Это отставание назовём *клином недоинвестирования*; его денежная цена равна вызванному им приросту совокупных издержек. Оптимум первого ранга восстанавливает внутренняя (пигувианская) цена информации:

$$p_{info} = \lambda\left|\frac{\partial \mathbb{E}[C_{loss}]}{\partial s}\right|_{s=s^*} = C'_{obs}(s^*). \tag{7}$$

Единица покрытия оценивается по экономии, которую она приносит фирме в рациональной точке, и бизнес-подразделения возмещают эксплуатации не учитываемую ею долю цены (7) (1 − κ); тогда решение центра затрат рационально для фирмы (ср. [2]). При κ = 0,3 покрытие снижается с 0,589 до 0,378, а совокупные издержки превышают рациональные на 26,2 %; внутренняя цена в калибровке равна 10,6 тыс. у.е. в год на процентный пункт покрытия.

Иллюстративная численная калибровка

Базовые значения параметров (расчёт на языке Python, начальное значение генератора 20260627, сценарий предоставляется по запросу): m = 1,0 у.е.; λ_{tx} = 1000 транзакций в минуту; $q_{free}(t) = q_{floor} + (1 - q_{floor})e^{-\theta_q t}$, q_{floor} = 0,05, θ_q = 0,05 мин$^{-1}$; $\alpha(t) = \alpha_{max}(1 - e^{-\psi t})$, α_{max} = 0,60, ψ = 0,03 мин$^{-1}$; γ = 0,50; δ = 0,30; θ = 0; пороги SLA (30; 60; 120) мин при штрафах (5000; 15 000; 40 000) у.е.; MTTR_{floor} = 8,0 мин, a = 52,0 мин, b = 2,2; MTTD той же формы: MTTD_{floor} = 1,5 мин, a_d = 18,0 мин, b_d = 2,8; k = 900 000 у.е./год, p = 1,8, λ = 60 инцидентов в год; $T = \mathrm{MTTR}\cdot\xi$, $\ln\xi \sim \mathcal{N}(-\zeta^2/2, \zeta^2)$, ζ = 0,35; σ ≡ 0.

На сетке до 180-й минуты удельная стоимость строго возрастает, а накопленные потери строго выпуклы (табл. 1, рис. 1). При σ ≡ 0 условие (D) выполняется автоматически, и расчёт позволяет оценить величину кривизны, но не проверяет само условие. Сумма $g_q + g_\Pi$ быстро убывает, поэтому в длительном инциденте или после раннего обнаружения сезонный спад спроса может нарушить (D).

Оптимум в калибровке внутренний (табл. 2, рис. 2). Рациональное покрытие сокращает совокупные издержки на 76,1 % по сравнению с нулевым, полное покрытие обходится на 55,7 % дороже рационального, а на канал обнаружения приходится 32,8 % предельной выгоды.

Сумма темпов опускается ниже 0,01 мин$^{-1}$ при $t \approx 43{,}8$ мин; у.е. — условные единицы

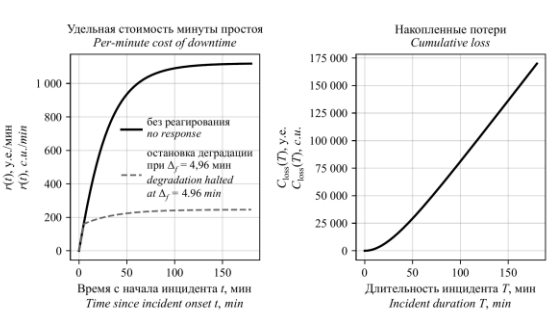


Рис. 1. Слева — r(t), справа — накопленные потери $C_{loss}(T)$ без реагирования и без учёта штрафов по SLA

Таблица 1
Удельная стоимость минуты простоя и темпы, входящие в условие (D)

Минута инцидента t *Incident minute t*	5	15	30	60
$r(t)$, у.е./мин / *c.u./min*	161,1	438,1	726,8	993,6
$r(t)/r(5)$	1,00	2,72	4,51	6,17
$g_q(t)$, мин$^{-1}$ / *min*$^{-1}$	0,1760	0,0448	0,0144	0,0026
$g_{\text{П}}(t)$, мин$^{-1}$ / *min*$^{-1}$	0,0162	0,0105	0,0059	0,0022
$g_q + g_{\text{П}}$, мин$^{-1}$ / *min*$^{-1}$	0,1922	0,0553	0,0203	0,0048

The sum of the rates falls below 0.01 *min*$^{-1}$ *at* $t \approx 43.8$ *min; c.u. stands for conventional units*

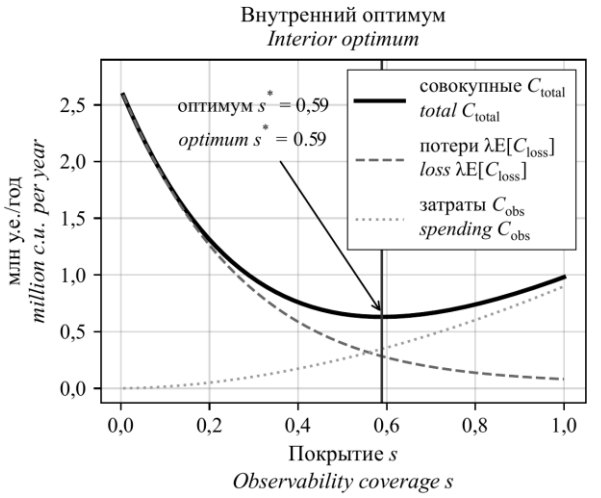


Рис. 2. Декомпозиция совокупных издержек $C_{\text{total}}(s) = \lambda\mathbb{E}[C_{\text{loss}}](s) + C_{\text{obs}}(s)$

Таблица 2
Базовые результаты иллюстративной калибровки

Показатель *Indicator*	Значение *Value*
Рациональное покрытие s^* *Rational coverage* s^*	0,589
MTTD при s^* / при $s \to 0$, мин MTTD *at* s^* */ at* $s \to 0$, *min*	4,96 / 19,5
MTTR при s^* / при $s \to 0$, мин MTTR *at* s^* */ at* $s \to 0$, *min*	22,2 / 60,0
C_{total} при $s \to 0$ / s^* / $s = 1$, млн у.е. C_{total} *at* $s \to 0$ / s^* / $s = 1$, *million c.u.*	2,633 / 0,630 / 0,981
Экономия при s^* / удорожание при $s = 1$ / доля C_{obs} в $C_{\text{total}}(s^*)$, % *Saving at* s^* */ extra cost at* $s = 1$ */ share of* C_{obs} *in* $C_{\text{total}}(s^*)$, *%*	76,1 / 55,7 / 55,1
Надбавка за минуту обнаружения λA / цена минуты устранения, тыс. у.е./год *Detection-minute premium* λA */ resolution-minute price, thousand c.u. per year*	35,9 / 22,8
Отношение полной ценности минуты обнаружения к цене минуты устранения; доля MB$_f$/MB, % *Full value of a detection minute relative to the resolution-minute price; share* MB$_f$/MB, *%*	2,58; 32,8
Эластичность $E(s^*)$; обе части правила (6) *Elasticity* $E(s^*)$*; both sides of rule* (6)	−0,461; 1,23

Однократные изменения параметров (шоки), указанные в примечании к табл. 3, подтверждают все знаки. В 300 испытаниях Монте-Карло по 13 параметрам ни один сдвиг не противоречил предсказанному (для SLA — расчётному) знаку; неподтверждёнными остались лишь сдвиги меньше допуска, в основном при полном или почти полном рациональном покрытии.

Таблица 3
Предсказания знаков в сравнительной статике и их проверка

Параметр *Parameter*	Знак *Sign*	Δs^*	Доля подтверждений *Share confirmed*	
			все, $N = 300$ *all,* $N = 300$	внутренние, $N = 228$ *interior,* $N = 228$
λ — частота инцидентов *Incident rate*	+	+0,083	0,763	1,000
m — маржа на транзакцию *Margin per transaction*	+	+0,065	0,763	1,000
Строгость SLA *SLA stringency*	+	+0,147	0,763	1,000
ψ — истощение терпения *Depletion of customer patience*	+	+0,008	0,750	0,982
γ — утрата пожизненной ценности *Loss of customer lifetime value*	+	+0,009	0,763	1,000
δ — возврат отложенного спроса *Return of deferred demand*	−	−0,015	0,787	0,996
θ — остаточный темп деградации; доля MB$_f$/MB *Residual degradation rate; detection share* MB$_f$/MB	−	32,8 → 10,3 %		

Шоки для столбца Δs^*: λ и m увеличены в 1,6 раза; ψ, γ, δ — в 1,5 раза; SLA: $T_k \cdot 0{,}6$ и $p_k \cdot 1{,}5$; θ: 0 → 0,5. В испытаниях Монте-Карло: λ, m, ψ увеличены в 1,5 раза; γ — на 0,5; δ — в 1,4 раза (не выше 0,95); SLA — как для столбца Δs^*. Допуск 0,001; внутренних решений ($0{,}01 < s^* < 0{,}99$) 228 из 300

Shocks for the Δs^* *column:* λ *and* m *multiplied by* 1.6*;* ψ, γ, δ *by* 1.5*; SLA:* $T_k \cdot 0.6$ *and* $p_k \cdot 1.5$*;* θ: 0 → 0.5. *In the Monte Carlo trials:* λ, m, ψ *multiplied by* 1.5*;* γ *increased by* 0.5*;* δ *multiplied by* 1.4*, capped at* 0.95*; SLA as above. Tolerance* 0.001*;* 228 *of* 300 *solutions are interior (*$0.01 < s^* < 0.99$*)*

Ограничения

Значения параметров заданы без оценивания, так что интерпретировать следует только знаки и соотношения, а условие (D) можно проверить лишь на производственной телеметрии. Недоучёт издержек ложных срабатываний смещает рациональное покрытие вверх, а допущение об экзогенной частоте инцидентов — вниз; учёт риска редких тяжёлых инцидентов повысил бы покрытие. Локализация первопричины [11] не заменяет эконометрической идентификации эффекта вмешательства.

Заключение

Получены четыре новых результата. Простой дорожает с каждой минутой, пока деградация и уход клиентов нарастают быстрее сезонного спада спроса, что проверяется по телеметрии и данным о клиентах. При углубляющейся деградации ускорение обнаружения ценнее равного ускорения устранения, так как раннее обнаружение её останавливает или замедляет. Эластичность времени восстановления по расходам определяется технологией, не зависит от уровня цен и меняется с покрытием, поэтому её оценки сопоставимы только при известном покрытии. Наконец, эксплуатация, не получающая всей выгоды, вкладывает меньше, чем выгодно фирме, и этот клин устраняет внутренняя цена информации, неучтённую эксплуатацией долю которой оплачивают бизнес-подразделения.

Для управления ИТ-компанией это означает, что бюджет наблюдаемости нужно определять сопоставлением предельной выгоды и предельных затрат, отказавшись от остаточного принципа. При рациональном бюджете рост расходов на 1 % снижает ожидаемый ущерб на столько процентов, сколько составляет отношение бюджета к ущербу; большее снижение говорит о заниженном бюджете, меньшее — о завышенном. Сбор всей телеметрии расточителен, когда затраты растут быстрее отдачи (в калибровке полное покрытие на 55,7 % дороже рационального), а вложения в обнаружение приоритетны, если деградация углубляется и её можно сдержать. Все числа получены в иллюстративной калибровке; следующим шагом станет эмпирическое оценивание модели на производственной телеметрии.

Литература

1. *Anderson R.* Why information security is hard – an economic perspective // Seventeenth Annual Computer Security Applications Conference. New Orleans, 2001. P. 358–365. DOI: 10.1109/ACSAC.2001.991552.
2. *Hirshleifer J.* On the economics of transfer pricing // The Journal of Business. 1956. Vol. 29. No. 3. P. 172–184. DOI: 10.1086/294110.
3. *Соколов Д. М.* Некоторые вопросы управления инцидентами в микросервисных архитектурах // Инновации и инвестиции. 2026. № 3. С. 443–447.
4. *Gordon L. A., Loeb M. P.* The economics of information security investment // ACM Transactions on Information and System Security. 2002. Vol. 5. No. 4. P. 438–457. DOI: 10.1145/581271.581274.
5. *Шуршев В. Ф., Гостюнин Ю. А.* Экспертная оценка ущерба при отказе информационной системы предприятия // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2020. № 4. С. 112–120. DOI: 10.24143/2072-9502-2020-4-112-120.
6. *Кайгородцев Г. И., Кравченко А. В.* Методика оценки эффективности информационных систем // Прикладная информатика. 2015. Т. 10. № 1(55). С. 5–14.
7. *Кацман В. Е.* Метод многоуровневой декомпозиции в экономических информационных системах // Прикладная информатика. 2006. № 3(3). С. 130–139.
8. *Barlow R., Hunter L.* Optimum preventive maintenance policies // Operations Research. 1960. Vol. 8. No. 1. P. 90–100. DOI: 10.1287/opre.8.1.90.

9. *Уринцов А. И.* Многоуровневые экономические информационные системы. – М.: Московский международный институт эконометрики, информатики, финансов и права, 2003. – 320 с.
10. *Li B., Peng X., Xiang Q., Wang H., Xie T., Sun J., Liu X.* Enjoy your observability: an industrial survey of microservice tracing and analysis // Empirical Software Engineering. 2022. Vol. 27. No. 1. Art. 25. DOI: 10.1007/s10664-021-10063-9.
11. *Xin R., Chen P., Zhao Z.* CausalRCA: Causal inference based precise fine-grained root cause localization for microservice applications // Journal of Systems and Software. 2023. Vol. 203. Art. 111724. DOI: 10.1016/j.jss.2023.111724.
12. *Garnett O., Mandelbaum A., Reiman M. I.* Designing a call center with impatient customers // Manufacturing & Service Operations Management. 2002. Vol. 4. No. 3. P. 208–227. DOI: 10.1287/msom.4.3.208.7753.
13. *Goldsmith-Pinkham P., Sorkin I., Swift H.* Bartik instruments: what, when, why, and how // American Economic Review. 2020. Vol. 110. No. 8. P. 2586–2624. DOI: 10.1257/aer.20181047.
14. *Heckman J. J.* Sample selection bias as a specification error // Econometrica. 1979. Vol. 47. No. 1. P. 153–161. DOI: 10.2307/1912352.

The economically rational level of observability spending in IT companies: marginal analysis and the internal price of information

Sokolov D.M.
Plekhanov Russian University of Economics

JEL classification: C01, C02, C1, C4, C5, C6, C8

The study is relevant because choosing an observability level means trading off loss-prevention spending against expected operational-risk losses. Costs fall on the operations unit and benefits on business units in other responsibility centres, so the level is set heuristically or residually and is systematically too low. The aim is to develop an economic and mathematical optimisation model of the observability-spending decision that links telemetry collection settings to monetary incident losses, and to determine the rational observability level and the conditions for reaching it within the firm. The problem is minimising expected incident losses plus observability spending over coverage. The methods are microeconomic demand modelling, marginal analysis, comparative statics, shadow prices, transfer pricing, an econometric identification design, numerical calibration and Monte Carlo simulation. The condition for a rising per-minute downtime cost reduces to an inequality testable from telemetry and customer data. Early detection is a separate benefit channel with its own shadow price, not implied by loss convexity. The spending elasticity of delays is independent of price levels and, given an irreducible minimum, non-monotonic in coverage; a budget-check rule in measurable quantities is derived. An internal price of information removes underinvestment. An illustrative calibration, which is not an estimate from production data, gives a rational coverage of 0.589, a cut in mean time to resolve (including detection) from 60.0 to 22.2 minutes and a 76.1% saving relative to zero coverage, while the detection channel yields 32.8% of the marginal benefit. Empirical estimation is the next step.

Keywords: economic and mathematical model, optimisation model, optimisation of decision-making, marginal analysis, observability spending, downtime cost, operational risk, elasticity, shadow price, transfer pricing, econometric identification, microservice architectures

References

1. Anderson R. Why information security is hard - an economic perspective. Seventeenth Annual Computer Security Applications Conference, New Orleans, 2001, pp.358-365. DOI: 10.1109/ACSAC.2001.991552.
2. Hirshleifer J. On the economics of transfer pricing. The Journal of Business, 1956, vol.29, no.3, pp.172-184. DOI: 10.1086/294110.
3. Sokolov D. M. Some Issues of Incident Management in Microservice Architectures. *Innovatsii i investitsii*, 2026, no.3, pp.443-447 (in Russian).
4. Gordon L. A., Loeb M. P. The economics of information security investment. ACM Transactions on Information and System Security, 2002, vol.5, no.4, pp.438-457. DOI: 10.1145/581271.581274.
5. Shurshev V. F., Gostiunin Iu. A. Expert evaluation of damage through enterprise information system failure. *Vestnik Astrakhanskogo gosudarstvennogo tekhnicheskogo universiteta. Seriya: Upravlenie, vychislitel'naya tekhnika i informatika*=Vestnik of Astrakhan State Technical University. Series: Management, Computer Science and Informatics, 2020, no.4, pp.112-120 (in Russian). DOI: 10.24143/2072-9502-2020-4-112-120.
6. Kaigorodtsev G. I., Kravchenko A. V. On the efficiency of information systems and the method of its estimation. *Prikladnaya informatika*=Journal of Applied Informatics, 2015, vol.10, no.1(55), pp.5-14 (in Russian).
7. Katsman V. E. *Metod mnogourovnevoi dekompozitsii v ekonomicheskikh informatsionnykh sistemakh* [The method of multilevel decomposition in economic information systems]. *Prikladnaya informatika*=Journal of Applied Informatics, 2006, no.3(3), pp.130-139.
8. Barlow R., Hunter L. Optimum preventive maintenance policies. Operations Research, 1960, vol.8, no.1, pp.90-100. DOI: 10.1287/opre.8.1.90.
9. Urintsov A. I. *Mnogourovnevye ekonomicheskie informatsionnye sistemy* [Multilevel economic information systems]. Moscow, *Moskovskii mezhdunarodnyi institut ekonometriki, informatiki, finansov i prava* Publ., 2003, 320 p.
10. Li B., Peng X., Xiang Q., Wang H., Xie T., Sun J., Liu X. Enjoy your observability: an industrial survey of microservice tracing and analysis. Empirical Software Engineering, 2022, vol.27, no.1, art.25. DOI: 10.1007/s10664-021-10063-9.
11. Xin R., Chen P., Zhao Z. CausalRCA: Causal inference based precise fine-grained root cause localization for microservice applications. Journal of Systems and Software, 2023, vol.203, art.111724. DOI: 10.1016/j.jss.2023.111724.
12. Garnett O., Mandelbaum A., Reiman M. I. Designing a call center with impatient customers. Manufacturing & Service Operations Management, 2002, vol.4, no.3, pp.208-227. DOI: 10.1287/msom.4.3.208.7753.
13. Goldsmith-Pinkham P., Sorkin I., Swift H. Bartik instruments: what, when, why, and how. American Economic Review, 2020, vol.110, no.8, pp.2586-2624. DOI: 10.1257/aer.20181047.
14. Heckman J. J. Sample selection bias as a specification error. Econometrica, 1979, vol.47, no.1, pp.153-161. DOI: 10.2307/1912352.