Interpretable Digital Traces: Mobile Wikipedia Usage and City-Level Tourism Demand

Lucas Eustache*
University Paris Dauphine – PSL

Paul Favier†
University Paris Dauphine – PSL

Work in progress — do not diffuse

Abstract

As tourism demand measurement increasingly relies on proprietary and costly data, freely accessible digital traces offer a scalable alternative. We propose mobile Wikipedia pageviews as an interpretable, high-frequency proxy for city-level hotel demand, grounded in a behavioral distinction between platforms: mobile browsing captures on-the-ground information seeking by physically present visitors, while desktop browsing reflects remote planning. Using a daily panel of 704 French cities (2018–2025) matched to proprietary hotel records from Accor, a two-way fixed-effects design controlling for city-level and daily aggregate shocks yields a robust positive elasticity between mobile Wikipedia traffic and same-day hotel room-nights. A COVID-19 natural experiment—exploiting the asymmetric collapse of mobile relative to desktop views during lockdowns—directly validates the physical-presence mechanism. Our central contribution is a systematic decomposition of this relationship at the city level, revealing that virtually all dispersion in destination-specific effects reflects true signal differences rather than estimation noise. The indicator performs best for internationally recognized tourist destinations and weakens—or reverses sign—for business-dominated and transit cities. A typological analysis identifies day-of-week desynchronization between Wikipedia activity and hotel demand as the main mechanism behind negative effects. Taken together, these findings map the boundary conditions under which mobile Wikipedia traffic constitutes a valid, cost-free tourism nowcasting instrument, with direct implications for destination monitoring and tourism policy.

Keywords: tourism demand; digital traces; Wikipedia; mobile information; high-frequency indicators; two-way fixed effects; treatment heterogeneity.
JEL: L83; C23; Z32; D83.

Introduction

Across markets, organizations increasingly rely on *digital trace data*—granular records of what people read, search, and click—to infer demand, monitor behavior, and allocate resources in near real time. Yet a persistent challenge is *interpretability*: the same online signal can reflect heterogeneous intentions (curiosity, learning, planning, or action), and thus its mapping to offline outcomes is often ambiguous. This challenge is particularly salient when managers and policymakers seek *high-frequency* indicators of local economic activity, but the most precise measures (e.g., mobile location data, transaction data) are typically proprietary, costly, and difficult to replicate. The question, therefore, is not merely whether online attention correlates with offline behavior, but whether we can extract *behaviorally meaningful*

*lucas.eustache@dauphine.psl.eu
†paul.favier@dauphine.psl.eu

structure from open digital traces that improves measurement, prediction, and theory.

Tourism provides a canonical setting where interpretability is both difficult and consequential. Travel decisions are experience-based; tourists face substantial information frictions and rely on ongoing search to reduce uncertainty before and during trips. At the same time, cities are policy-relevant units that can invest heavily in amenities, branding, and information infrastructures, yet they often lack accessible indicators of visitation. Traditional tourism statistics are frequently released with delays, while proprietary alternatives—such as mobile location data or payment records—are costly, opaque, and unavailable to most stakeholders. Episodic shocks from transport disruptions to public health restrictions further underscore the need for *transparent, replicable* measurement tools. Recent research shows that platform information and user-generated content shape travel-related choice and market outcomes. Still, we know less about whether the *composition* of online information consumption contains incremental signal about contemporaneous, on-the-ground tourism activity—and whether such signals can be extracted from publicly available data.

This paper argues that *where* and *how* information is accessed is central to that signal. Mobile devices embed information seeking in time and space: smartphones reduce the friction of "in-the-moment" queries but impose interface constraints (e.g., smaller screens) that alter browsing costs and depth. In contrast, desktop access is more consistent with remote, leisurely browsing and trip planning. Prior work in information systems and marketing demonstrates that mobile contexts heighten locality and immediacy, changing the salience of local activities and the economics of search. In tourism specifically, smartphones support navigation, re-planning, and coordination during travel, and tourists' information needs shift across inspiration, planning, and on-site execution. These mechanisms imply that device-based differences in information consumption can serve as a theoretically grounded proxy for the timing and context of offline activity.

We study this idea using Wikipedia, a general-purpose, non-transactional information infrastructure whose content is publicly accessible and widely consulted. Unlike vertically integrated travel platforms, Wikipedia is not optimized for conversion; it functions as a neutral reference layer that travelers can use to resolve factual uncertainty and orient themselves at a destination. Moreover, a growing literature shows that Wikipedia's governance and content dynamics shape information quality and bias, and causal evidence indicates that improving city-page content can increase tourism outcomes. These features make Wikipedia particularly useful for disentangling *information consumption as behavior* from platform-mediated transactions.

Our core measurement insight is that *device composition* can sharpen the behavioral interpretation of online attention. Total pageviews of a city's Wikipedia page may reflect diffuse salience (e.g., education, news, general curiosity) and thus provide a noisy proxy for visitation. By contrast, the *share of mobile pageviews* plausibly captures the fraction of attention generated in "on-the-go" contexts that align more tightly with contemporaneous physical presence. This yields our two research questions:

> **RQ1**: Do daily mobile Wikipedia pageviews serve as a high-frequency proxy for same-day hotel room-night demand at the city level?
>
> **RQ2**: How does the Wikipedia–tourism relationship vary across cities, and what city characteristics explain this heterogeneity?

We answer these questions using a daily city-level panel linking Accor hotel data to Wikipedia mobile pageviews for 704 French cities over January 2018–June 2025 (approximately 1.6 million city-day observations). Leveraging within-city, within-date variation through two-way fixed effects, we find a global

semi-elasticity of +0.125. We then apply the Frisch-Waugh-Lovell theorem to estimate city-specific coefficients and document strong, predictable heterogeneity—with an I^2 of 98.0% and a fourfold gradient across Wikipedia visibility quintiles.

This study contributes to the literature in four ways. First, it advances research on digital traces by demonstrating that *daily-frequency* mobile Wikipedia data can serve as a real-time tourism indicator, extending prior work that relies on monthly or annual aggregates. Second, it addresses the interpretability problem by showing that *composition-based* measures (mobile versus desktop) improve signal quality through a device-specific behavioral mechanism. Third, it provides the first systematic analysis of *city-level heterogeneity* in the Wikipedia–tourism relationship, documenting its magnitude ($I^2 = 98.0\%$), its determinants (visibility gradient, tourism type), and its failure modes (business-travel cities, excursionist destinations). Fourth, it characterizes the subset of cities for which the Wikipedia proxy is unreliable—negative-effect cities—by identifying temporal desynchronization mechanisms (day-of-week mismatch, seasonal mismatch) that explain why the proxy fails in those contexts. For destination managers and policymakers, the results imply that mobile Wikipedia monitoring is highly effective for established tourist destinations and should be calibrated—or supplemented—for business-dominated or excursionist cities.

1 Literature Review

This paper builds on research using digital traces to measure economic activity and on work that explains when information consumption maps into offline behavior. The interpretability problem is central: online attention can reflect heterogeneous motives, so identifying trace features that correspond to specific behavioral contexts is key. Our focus is on tourism, where prior work shows that online information environments and user-generated content shape travel-related choice and firm outcomes (Ghose, Ipeirotis, et al. (2012); Hollenbeck et al. (2019); Gao et al. (2025)). We extend this line of inquiry by asking whether *device composition* (the share of mobile versus desktop consumption) contains incremental signal for *contemporaneous* in-destination tourism activity.

Digital traces as proxies for offline demand

A broad literature links online traces to offline outcomes by treating digital activity as observable manifestations of attention, word-of-mouth, and diffusion. Online text and interaction data have been used as measurable proxies for these constructs, with links to demand-side outcomes such as sales and media consumption (Dellarocas (2003); Godes et al. (2004); Chevalier et al. (2006)). The underlying premise is that online activity reflects information acquisition and social influence processes that map into economic behavior: online conversations predict TV ratings in dynamic demand models (Godes et al. (2004)), and review valence and volume relate to relative sales across retailers (Chevalier et al. (2006)).

In travel markets, platform information and user generated content (UGC) influence both consumer and firm behavior. Crowd-sourced hotel content can improve rankings and better match multi-attribute preferences (Ghose, Ipeirotis, et al. (2012)), ratings shape hotels' advertising responses (Hollenbeck et al. (2019)), and interventions that alter review generation produce downstream marketplace effects (Gao et al. (2025)). Together, this stream supports the plausibility that online information use co-moves with contemporaneous activity, especially in experience-good settings where information reduces uncertainty. At the same time, these studies typically emphasize the *level* of online activity or content, leaving open whether trace *composition* helps distinguish contexts (e.g., armchair interest versus on-site need) that matter for real-time measurement.

Search frictions

The micro-foundations for why online attention should relate to visitation come from consumer search under information frictions. Economics documents costly search, choice-set formation, and the use of browsing data to test search models and infer frictions (Bakos (1997); Brynjolfsson et al. (2000); De Los Santos et al. (2012)). Consumer research similarly emphasizes constructive preferences and context-dependent heuristics that shape what information is processed and how alternatives are screened (Bettman et al. (1998); Häubl et al. (2000)). These perspectives jointly imply that information consumption can track behavior most closely when uncertainty is salient and decisions are time-sensitive.

Tourism is a particularly information-intensive setting because information needs unfold across inspiration, planning, and on-site execution, and smartphones expand tourists' ability to search, re-plan, and coordinate during the trip (Dickinson et al. (2016)). This makes a "same-day" link theoretically coherent: if tourists consult contextual facts, navigation cues, attractions, and local services while traveling, online information consumption should rise during physical presence at the destination (Huertas et al. (2022)). The remaining challenge is to identify which trace features correspond to on-site information use rather than broader salience that may not translate into visits.

Mobile context, device constraints, and composition-based signals

This is where mobile versus desktop becomes consequential. Device constraints and usage contexts change search costs and browsing depth. Ghose, Goldfarb, et al. (2013) show that smaller screens raise browsing costs on mobile, while mobile usage across offline locations increases the salience of local activities, implying systematic differences in what users seek on mobile versus PC. Marketing research likewise emphasizes that mobile embeds consumers in time and space, so proximity and immediacy shape responsiveness to information and persuasion (Fong et al. (2015)). In tourism, these mechanisms align with evidence that tourists search at the destination with topic- and situation-specific patterns (Huertas et al. (2022)) and that smartphones support logistical and coordination needs during travel (Dickinson et al. (2016)).

Applied to Wikipedia pageviews, device composition is informative because it shifts attention from *how much* is consumed to *where and when* consumption likely occurs. Our empirical object is not "Wikipedia attention predicts future tourism," but whether the device composition of Wikipedia readership tracks contemporaneous tourist presence. Total pageviews can reflect diffuse interest (students, remote curiosity, national news), but a higher mobile share plausibly indicates a higher fraction of users consuming the page while traveling. Because the measure is a share, it can contain incremental signal beyond levels: holding total attention fixed, a shift from desktop to mobile may indicate a transition from low-intent or armchair consumption to situated, context-dependent consumption that should map more tightly into same-day visitation (Ghose, Goldfarb, et al. (2013)).

This mechanism also highlights boundary conditions. First, seasonality and destination heterogeneity matter: tourism is strongly seasonal, and cities vary in baseline "touristicity" so the same device mix may translate differently into visitation across high-capacity destinations versus places where attention is primarily non-touristic. Second, the information environment quality can mediate the attention–presence link: richer, better-structured information improves screening and choice (Häubl et al. (2000)), while Wikipedia pages vary in depth and stability across space and time (Aaltonen et al. (2016); Ransbotham et al. (2011)). Thus, mobile attention may track visitation more strongly when pages are information-rich and stable, and more weakly when pages are sparse or low-quality.

Wikipedia as an open information infrastructure

Wikipedia is a relevant platform for this mechanism because it functions as general-purpose information infrastructure rather than a vertically integrated travel marketplace. Its institutional logic is oriented toward neutral, encyclopedic coverage, and governance and production research shows that neutrality and bias depend on contributor composition and coordination (Ransbotham et al. (2011); Greenstein et al. (2016); Greenstein et al. (2018)), while content supply follows systematic dynamics that shape what information is available and when (Aaltonen et al. (2016)). Wikipedia content can also affect real outcomes: a randomized field experiment adding tourist-relevant content to city pages increased overnight stays, with effects operating through increased Wikipedia readership (Hinnosaar et al. (2023)). Consistent with this intermediary role, statistical agencies and applied work have explored Wikipedia pageviews as proxies for visitation intensity and seasonality patterns (Owuor et al. (2023)).

Hypotheses

Taken together, existing research establishes (i) that digital traces can proxy offline demand, (ii) that search frictions and context shape when information use coincides with behavior, and (iii) that mobile devices systematically alter the economics and locality of search. However, the literature provides less direct evidence on whether *device composition* of a general-purpose information infrastructure can be used as a *contemporaneous, daily-frequency* indicator of local tourism activity at the city level, above and beyond overall attention and content supply. This motivates our research questions, which we operationalize through the following hypotheses.

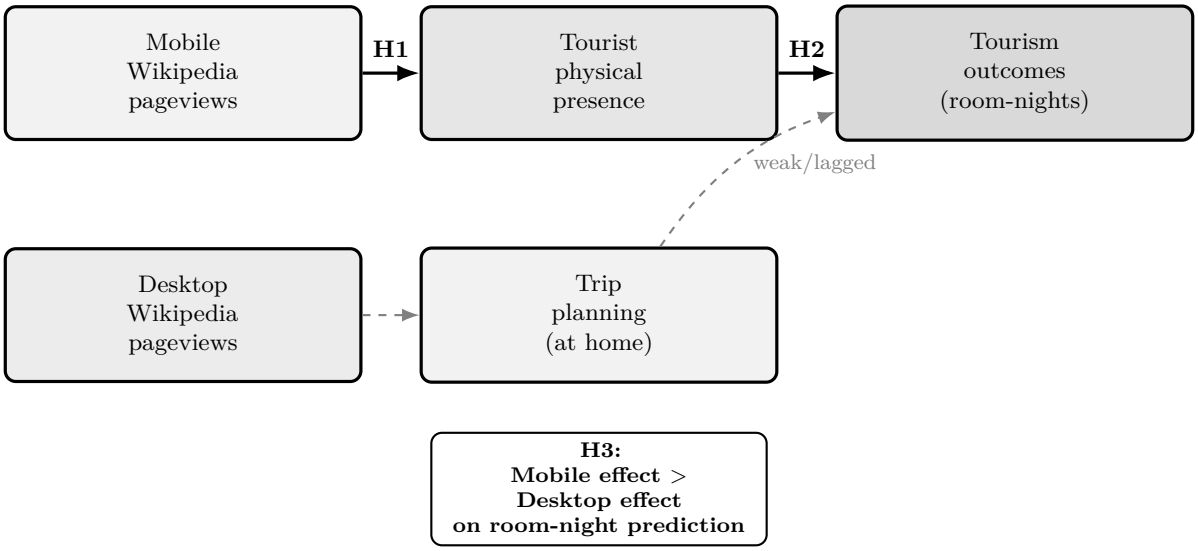

Figure 1: Conceptual model: Mobile Wikipedia pageviews as a contemporaneous proxy for tourism activity.

Hypothesis 1: Mobile pageviews reflect physical presence. When tourists are on site, they generate mobile Wikipedia pageviews while satisfying immediate informational needs. As a result, mobile pageviews for a city's Wikipedia page should move contemporaneously with actual tourist presence in that city.

Hypothesis 2: Physical presence generates contemporaneous hotel demand. Tourist visits—the same episodes that generate mobile Wikipedia pageviews (H1)—produce hotel room-nights. Physical presence is thus the common cause of both signals: it simultaneously drives mobile browsing and

accommodation demand. Consequently, mobile pageviews and room-nights should co-vary within the same city on the same day: $\beta_{\text{mobile}} > 0$ in a regression of log room-nights on log mobile pageviews with city and date fixed effects.

Hypothesis 3: Mobile effect exceeds desktop effect. Because mobile pageviews are tightly synchronized with physical presence while desktop pageviews capture more temporally diffuse planning behavior, the mobile coefficient should exceed the desktop coefficient: $\beta_{\text{mobile}} > \beta_{\text{desktop}}$.

Hypothesis 4: Heterogeneity of effect. The effect of Wikipedia mobile pageviews on hotel demand should be stronger in settings where informational needs are higher (established tourist destinations with rich Wikipedia content) and weaker or even absent in settings where Wikipedia attention reflects non-touristic motives (business-travel cities, excursionist destinations).

2 Data

Our empirical analysis rests on three sources fused at the city-day level: (i) proprietary hotel-performance records from Accor Group, (ii) daily Wikipedia pageview data from the Wikimedia REST API, and (iii) municipal typology metadata from INSEE and a clustering exercise. The analytical sample consists of **1,635,454 city-day observations** covering **704 French communes** over **2,738 days** (1 January 2018 – 30 June 2025). Table 1 details the successive construction steps from raw proprietary sources to the final sample.

Table 1: Data construction pipeline: from raw sources to analytical sample

Step	Description	Observations	Communes
1a	Wikipedia raw (daily city pageviews, Wikimedia API)	1,927,552	704
1b	Accor raw (reservation records, all channels)	5,900,336	707
2	Accor: pivot from long to city-day format	1,680,331	707
3	Inner join Wikipedia × Accor on (INSEE code, date)	1,640,595	704[a]
4	Quality filter: drop city-days with zero total pageviews	1,635,454	704
5	FWL subsample (≥200 obs, ≥100 days per city)	—	697

[a] Three communes excluded because their Accor records begin only after the Wikipedia extraction ends (July 2025): Barjouville (28024), Carbon-Blanc (33096), L'Isle-sur-la-Sorgue (84054).

2.1 Tourism Data

The hotel-performance data are provided by Accor Group, the leading hotel operator in France, under a proprietary data-sharing agreement. The raw file records 5,900,336 reservation lines covering 753 establishment names (mapped to 707 unique INSEE codes via a city reference table), spanning January 2018 to August 2025, and disaggregated by six distribution channels (*Web Direct*, *Web Indirect*, *Intra-TARS*, *GDS/IDS*[1], *Call Center*, *Unknown*). We pivot these records to a city-day format, summing room-nights across all channels to obtain $ROOM\_NIGHT\_TOTAL_{it}$.

Summation is the semantically appropriate aggregation for room-nights: the total is an additive volume measure, so collapsing across channels and establishments by sum is well-defined.[2] Our primary dependent variable is:

[1] Global Distribution Systems (GDS) is a travel service reservation management software.

[2] By contrast, the raw file also provides *AVG_OCCUPATION_RATE*, which cannot be used directly: the pivot script aggregates it by sum rather than as a weighted average, producing values exceeding 100% for cities with multiple Accor properties (5.3% of observations). Additionally, *STAY_BEGINNING_DATE*, the only date field in the Accor source, is considered as a daily occupancy snapshot, which is relevant to measure same-day hotel activity.

$$\text{LogRN}_{it} = \log(1 + \text{ROOM\_NIGHT\_TOTAL}_{it}),$$

the log-transformation of daily room-nights. The raw variable is strongly right-skewed (mean: 93.9; median: 35.0; skewness: 23.9), driven by large cities and peak-season days. The log transformation reduces skewness to 0.29, bringing the distribution close to symmetry and enabling a semi-elasticity interpretation of regression coefficients.

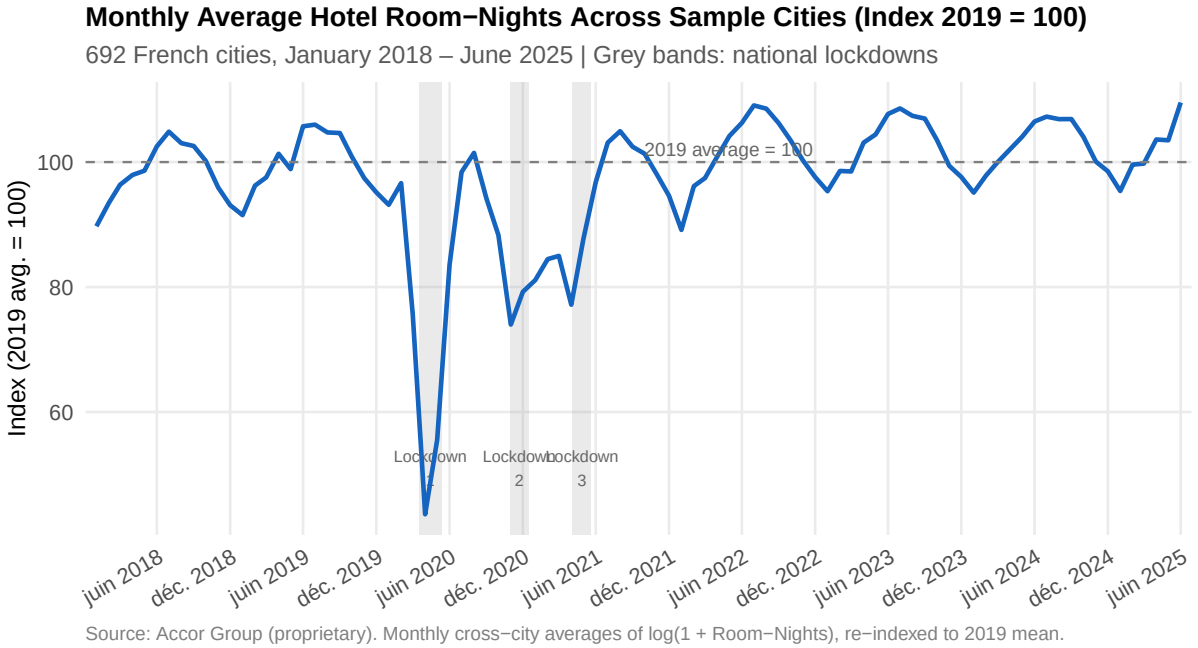

Figure 2: Monthly average hotel room-nights across sample cities, indexed to 2019 (= 100). *Note:* Grey bands mark the three French national lockdown periods (March–May 2020, October–December 2020, April–May 2021). Index computed from cross-city averages of log(1 + Room-Nights), re-scaled to the 2019 mean. Source: Accor Group (proprietary data).

Figure 2 plots the average monthly room-night volume across the 704 sample cities, indexed to the 2019 mean. The series displays strong seasonal patterns, with peaks in July–August and secondary peaks around school holidays, and troughs in January–February. The sharp decline in early 2020 reflects the COVID-19 shock and the successive national lockdowns, followed by a gradual recovery to above-2019 levels by summer 2022.

2.2 Wikipedia Data

For each city, we retrieve daily pageview counts from the Wikimedia REST API.[3] The API distinguishes human traffic from automated agents (bots and spiders) and separates access platforms (desktop, mobile-web, mobile-app). Our primary attention variable uses **mobile-web pageviews by human users** exclusively (`pv_mobile_web_user`); mobile-app traffic is excluded because its daily volume is negligible (mean: 3.3 views per city-day, versus 109.7 for mobile-web) and its on-the-ground interpretation is less clean. We construct three measures:

[3]Appendix B details the Wikipedia data extraction procedure.

$$\text{MobileLog}_{it} = \log(1 + \text{MobileWebViews}_{it}), \tag{1}$$

$$\text{DesktopLog}_{it} = \log(1 + \text{DesktopViews}_{it}), \tag{2}$$

$$\text{MobileShare}_{it} = \frac{\text{MobileWebViews}_{it}}{\text{AllAccessViews}_{it}} \times 100, \tag{3}$$

where *AllAccessViews* aggregates desktop, mobile-web and mobile-app human traffic. The mean daily mobile-web views across qualifying city-days is 109.7 (median: 54; SD: 193.2), with substantial right-skew driven by major tourist cities and occasional viral events. The average mobile share is 58.5% in the daily panel, reflecting the dominance of smartphone usage throughout the observation period. The quality filter in Step 4 (Table 1) drops city-days with zero total pageviews—i.e. days for which the Wikipedia API returned no activity—removing 5,141 observations (0.3%); all 704 communes are retained, as no city has zero total pageviews across its entire observation window.

To characterize tourism-specific content, we parse each page's HTML monthly and extract section-level measures for tourism-relevant parts (sections with headings matching keywords like "Tourisme," "Lieux et monuments," "Patrimoine"). In the sample, the tourism section averages 1,403 words (SD: 1,609) and 9.4 images (SD: 11.8). We also use the number of Wikipedia language versions for each city (mean: 78) as a proxy for international notoriety.

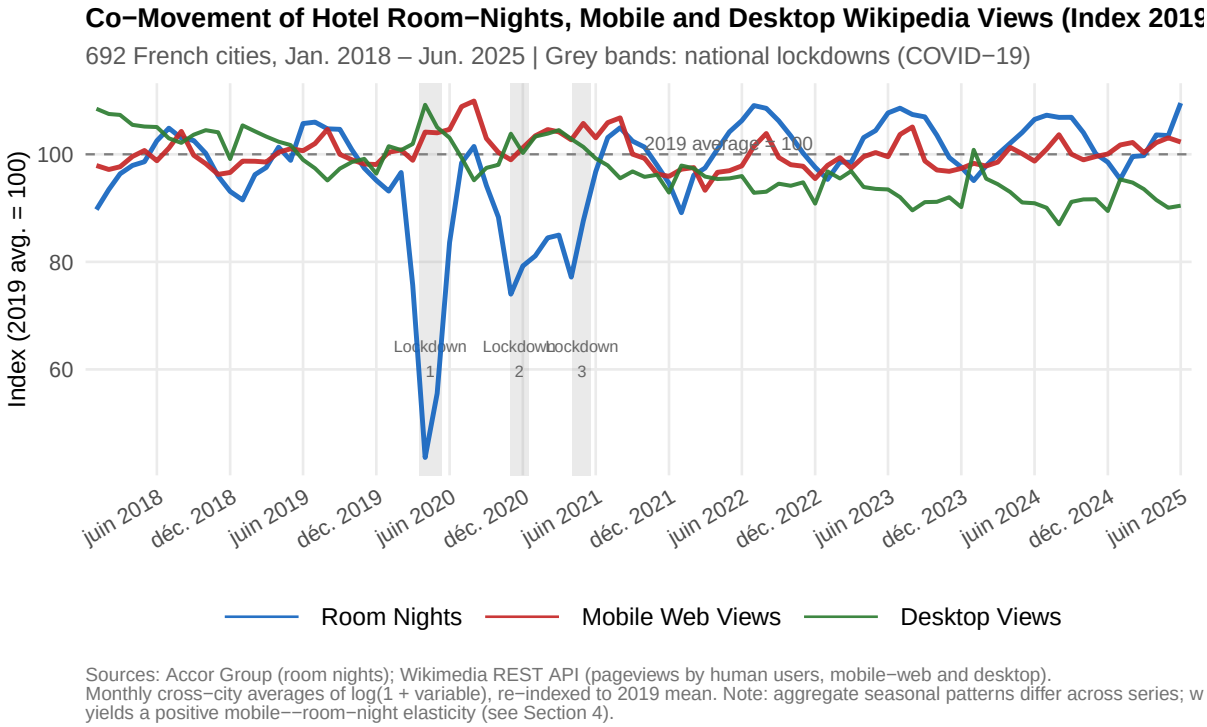

Figure 3: Co-movement of hotel room-nights, mobile Wikipedia views, and desktop Wikipedia views, indexed to 2019 (= 100). *Note:* Grey bands mark the three French national lockdown periods. All three series are monthly cross-city averages of log(1 + variable), re-scaled to the 2019 mean. Mobile and desktop series use human-generated traffic only (`pv_mobile_web_user` and `pv_desktop_user`). Sources: Accor Group; Wikimedia REST API.

Figure 3 displays the three series on a common index. Two patterns are noteworthy. First, all three variables collapsed during the 2020 lockdowns and recovered thereafter, consistent with the physical-

presence mechanism: hotel demand, mobile browsing, and even desktop browsing all contracted sharply when mobility was restricted. Second, the aggregate seasonal profiles differ across platforms—room-nights peak strongly in summer (July–August), while Wikipedia traffic shows a flatter seasonal pattern with relatively higher activity in autumn and winter, reflecting trip-planning and general-curiosity browsing outside peak tourism season. This divergence in aggregate dynamics illustrates why a naïve time-series correlation between Wikipedia traffic and hotel demand would be confounded by seasonal composition effects, and motivates the two-way fixed-effects design: identification exploits within-city, within-day deviations from city and date averages rather than the aggregate co-movement visible here.

2.3 Supplementary Metadata

We merge two supplementary datasets to support the heterogeneity analysis. First, official city-level statistics from INSEE provide baseline controls: resident population, accommodation supply (numbers of hotels, campsites, and other lodging establishments), and tourism demand per department. Second, a pre-constructed city-level typology—derived from a separate clustering exercise—provides binary indicators for ski resorts (`is_ski`), seaside destinations (`is_seaside`), UNESCO-listed sites (`is_unesco`), and cities holding a tourism label (`has_label`), as well as continuous measures of Wikipedia internationalization (`lang_count`), hotel capacity (`rooms_mean`), and information density (`info_density`). These typological variables are used exclusively in the heterogeneity analysis (Section 4.4); the main identification relies on city and date fixed effects only.

3 Methodology

Our empirical strategy exploits within-city variation over time to assess whether mobile Wikipedia pageviews proxy for contemporaneous hotel demand. We present a primary specification using the daily panel, supplemented by the monthly aggregation for robustness and comparison.

3.1 Primary Specification: Two-Way Fixed Effects (Daily Panel)

The core identifying premise is a behavioral distinction across access platforms. Mobile Wikipedia pageviews primarily reflect in-situ, "on-the-ground" information seeking by visitors physically present at a destination, whereas desktop pageviews more strongly capture remote search and trip planning.

City fixed effects (α_i) absorb time-invariant heterogeneity (baseline attractiveness, tourism infrastructure, population size). Date fixed effects (λ_t) capture aggregate shocks common to all cities on a given day: national and regional holidays, day-of-week effects, seasonal patterns, macroeconomic conditions, and COVID-19 lockdowns. Identification relies on the question: on days when a city receives unusually high mobile Wikipedia traffic *relative to the national average for that day*, does it also register unusually high hotel room-nights?

The primary specification is:

$$\log(1 + \text{RN})_{it} = \beta \cdot \text{MobileLog}_{it} + \alpha_i + \lambda_t + \varepsilon_{it}, \tag{4}$$

where i indexes cities, t indexes days, and standard errors are clustered at the city level. We implement four principal variants:

1. Baseline: $X_{it} = \text{MobileLog}_{it}$
2. Desktop: $X_{it} = \text{DesktopLog}_{it}$

3. Mobile share: $X_{it} = \text{MobileShare}_{it}$

4. Joint: $X_{it} = (\text{MobileLog}_{it}, \text{DesktopLog}_{it})$

Day-of-week confounding. An important identification concern unique to the daily panel is day-of-week (DOW) confounding. Mobile Wikipedia views exhibit systematic weekly patterns (higher on weekends, consistent with leisure browsing), as do hotel room-nights (which vary by destination type). Because our date fixed effects are defined at the daily level—each of the 2,738 days in the sample receives its own fixed effect—the DOW pattern is automatically absorbed into λ_t. This is a key advantage over specifications with coarser time fixed effects (e.g., month-year dummies), where DOW remains a potential confounder. We document this empirically in Section 4.4.1.

3.2 Robustness: Monthly Aggregation

To assess robustness to temporal aggregation, we collapse the daily city-day panel to the city-month level and re-estimate the same specification:

$$\log(1 + \text{RN})_{im} = \beta \cdot X_{im} + \alpha_i + \lambda_m + \varepsilon_{im}, \tag{5}$$

at the city-month level, where $\log(1 + \text{RN})_{im}$ is monthly log room-nights (summed over days within the month), λ_m are month-year fixed effects, and standard errors are clustered at the city level.

3.3 City-Level Heterogeneity: FWL Decomposition

To estimate city-specific effects without the computational burden of interacting city dummies with the main variable, we apply the Frisch-Waugh-Lovell (FWL) theorem. We first partial out the date fixed effects by computing date-demeaned residuals for the entire panel:

$$\widetilde{\text{MobileLog}}_{it} = \text{MobileLog}_{it} - \overline{\text{MobileLog}}_{\cdot t}, \tag{6}$$

$$\widetilde{\text{LogRN}}_{it} = \text{LogRN}_{it} - \overline{\text{LogRN}}_{\cdot t}, \tag{7}$$

where $\bar{\cdot}_t$ denotes the cross-city mean on day t. By FWL, the global two-way FE estimate $\hat{\beta}$ is algebraically equivalent to the OLS estimate from regressing $\widetilde{\text{LogRN}}$ on $\widetilde{\text{MobileLog}}$ with city fixed effects. City-specific estimates are then obtained by running a separate OLS on the residuals for each city:

$$\hat{\beta}_i = \frac{\sum_t \widetilde{\text{MobileLog}}_{it} \cdot \widetilde{\text{LogRN}}_{it}}{\sum_t \widetilde{\text{MobileLog}}_{it}^2}. \tag{8}$$

These city-specific coefficients $\hat{\beta}_i$ use exactly the same source of variation as the global 2FE—the cross-sectional temporal deviation "which city has relatively more mobile views than others on day t?"—but allow for full heterogeneity in the effect size. By construction, the variance-weighted average of $\hat{\beta}_i$ recovers the global $\hat{\beta}$, providing an internal consistency check. We qualify cities for estimation by requiring at least 200 daily observations, 100 distinct dates, and positive variance in the residualized mobile views.

Formal heterogeneity test. We assess whether the dispersion in $\hat{\beta}_i$ exceeds what would be expected from sampling noise using the Cochran Q statistic and its derived I^2 index:

$$Q = \sum_{i=1}^{k} w_i(\hat{\beta}_i - \hat{\beta}_{\text{pooled}})^2, \quad w_i = 1/\hat{s}_i^2, \tag{9}$$

where $\hat{\beta}_{\text{pooled}} = \sum w_i \hat{\beta}_i / \sum w_i$ and $\hat{s}_i$ is the heteroskedasticity-robust standard error of $\hat{\beta}_i$. The I^2 statistic, defined as $\max(0, (Q - (k-1))/Q) \times 100$, measures the percentage of total variation attributable to true heterogeneity rather than sampling error.

3.4 COVID-19 as a Mobility Shock

The COVID-19 pandemic provides a large, plausibly exogenous shock to physical mobility. If mobile pageviews predominantly reflect on-site behavior, they should decline disproportionately relative to desktop pageviews during lockdowns. We implement a difference-in-differences analysis comparing mobile versus desktop pageviews (treatment dimension) across lockdown versus non-lockdown periods (time dimension):

$$\log(1 + \text{Views}_{ipt}) = \gamma_1 \text{Mobile}_p + \gamma_2 \text{Lockdown}_t + \gamma_3 (\text{Mobile}_p \times \text{Lockdown}_t) + \alpha_i + \varepsilon_{ipt}, \tag{10}$$

where $p \in \{\text{mobile}, \text{desktop}\}$ and Lockdown_t flags March–May 2020 (first French national lockdown) and November–December 2020 (second lockdown). The interaction $\gamma_3 < 0$ would confirm that mobile views fall more than desktop during lockdowns, consistent with the on-the-ground mechanism.

4 Results

4.1 Main Results: Daily Panel

Table 2 reports estimates from the daily two-way fixed-effects model. The dependent variable is $\log(1 + \text{Room-nights})$. All specifications include city fixed effects and date fixed effects, with standard errors clustered by city.

Table 2: Mobile Wikipedia pageviews and daily hotel room-nights: Primary results

	(1) log(Mobile)	(2) log(Desktop)	(3) Mobile share	(4) Both
log(1+Mobile views)	0.125***			0.116***
	(0.009)			(0.008)
log(1+Desktop views)		0.067***		0.035***
		(0.006)		(0.004)
Mobile share (%)			0.001***	
			(0.000)	
City FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Observations	1,635,454	1,635,454	1,635,454	1,635,454
Within R^2	0.007	0.002	0.001	0.007

Standard errors clustered by city in parentheses.
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Main finding. Column (1) shows that a one-unit increase in $\log(1 + \text{Mobile views})$ is associated with a 0.125-unit increase in $\log(1 + \text{Room-nights})$ within the same city on the same day ($p < 0.001$). This

coefficient is a semi-elasticity: holding all date-specific shocks constant, a city that attracts 10% more mobile Wikipedia traffic than its daily average registers approximately 1.25% more hotel room-nights. The within R^2 of 0.7% is modest, consistent with the many determinants of hotel demand beyond Wikipedia browsing.

Mobile vs. desktop decomposition. Columns (2)–(4) decompose total pageviews by platform. Desktop pageviews alone also display a positive association with room-nights ($\beta = 0.067$, within $R^2 = 0.002$), though the coefficient is roughly half the size of the mobile estimate. When both variables are included jointly (column 4), the mobile coefficient stabilises ($\beta = 0.116$) while the desktop coefficient is substantially attenuated ($\beta = 0.035$), indicating that mobile browsing captures most of the identifying variation. A Wald test rejects coefficient equality ($H_0 : \beta_{\text{mobile}} = \beta_{\text{desktop}}$) at $p < 0.001$, consistent with Hypothesis 3.

Robustness: Monthly aggregation. Table 3 shows estimates from the city-month panel, using log(1 + Monthly room-nights) as the dependent variable. The mobile coefficient ($\hat{\beta} = 0.372$, $p < 0.001$) is positive and significant across all specifications. When both platforms are included jointly (column 4), the mobile coefficient remains large ($\hat{\beta} = 0.373$) while the desktop coefficient becomes negligible ($\hat{\beta} = -0.002$), mirroring the platform-dominance pattern observed in the daily panel. These monthly-aggregated results confirm that the association is not an artifact of the daily temporal resolution.

Table 3: Wikipedia pageviews and hotel room-nights: Monthly aggregation (robustness)

	(1) Mobile share	(2) log(Mobile)	(3) log(Desktop)	(4) Both
Mobile share (%)	0.010***			
	(0.002)			
log(1+Mobile views)		0.372***		0.373***
		(0.028)		(0.034)
log(1+Desktop views)			0.283***	−0.002
			(0.051)	(0.047)
City FE	Yes	Yes	Yes	Yes
Month-Year FE	Yes	Yes	Yes	Yes
Observations	56,024	56,024	56,024	56,024
Within R^2	0.007	0.024	0.008	0.024

Dep. var.: log(1 + Monthly Room-Nights) (monthly sum). Standard errors clustered by city.
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

4.2 Leads-and-Lags Analysis

A key identification concern is reverse causality: tourism activity may drive Wikipedia pageviews rather than the converse. We address this by estimating eleven separate models with leads and lags of the mobile variable spanning a ±60-day horizon (−60, −30, −14, −7, −1, 0, +1, +7, +14, +30, +60 days). All models are estimated on the same restricted sample (communes with at least 60 observations on each side), enabling direct coefficient comparison. Table 4 reports a symmetric selection of key horizons; the full eleven-point results appear in Appendix C.4.

The results display a pronounced bell-shaped pattern centred on t and $t + 1$, with coefficients declining monotonically as the horizon increases in either direction. The contemporaneous effect ($\hat{\beta} = 0.125$) and one-day lead ($\hat{\beta} = 0.130$) are the two largest estimates; the $t + 1$ value marginally exceeds the contemporaneous one, consistent with tourists consulting Wikipedia on the evening of or the day after

Table 4: Placebo test: Leads and lags of log mobile pageviews (daily panel, ±60-day window, selected horizons)

Timing	$\hat{\beta}$	SE	95% CI	p-value
$t-60$ (lag 60 days)	0.023***	0.004	[0.015, 0.032]	< 0.001
$t-14$ (lag 14 days)	0.093***	0.007	[0.079, 0.107]	< 0.001
$t-1$ (lag 1 day)	0.110***	0.008	[0.094, 0.125]	< 0.001
t (contemporaneous)	0.125***	0.009	[0.108, 0.143]	< 0.001
$t+1$ (lead 1 day)	0.130***	0.009	[0.113, 0.148]	< 0.001
$t+14$ (lead 14 days)	0.086***	0.007	[0.073, 0.099]	< 0.001
$t+60$ (lead 60 days)	0.015***	0.004	[0.006, 0.023]	< 0.001

*** $p < 0.001$. Daily panel; dep. var.: log(1 + Room-nights).
Each row is a separate regression with city and date FE, SE clustered by city.
All models on the same sample (restricted to ±60-day margin). Full results in Appendix C.4.

arrival. Moving to the 14-day horizons, the coefficient falls to roughly 0.093–0.086, and by 60 days it shrinks to 0.023–0.015—less than one-fifth of the peak and only marginally significant. The near-zero effect at ±60 days is inconsistent with pure reverse causality or with a spurious common trend, both of which would predict large and persistent lead effects. The approximate symmetry of the lag and lead profiles further suggests that the dominant channel is contemporaneous co-movement driven by on-the-ground tourist presence, rather than anticipation or post-visit search. Full regression output is in Appendix C.4.

[WIP] *However, the fact that both the 60-day lag and lead remain statistically significant—albeit small—raises a residual concern: both Wikipedia pageviews and hotel room-nights are daily time series exhibiting strong autocorrelation (seasonal cycles, weekly patterns, trend). Even after absorbing city and date fixed effects, residual serial correlation may sustain non-zero cross-correlations at long horizons that do not reflect any direct causal link. In other words, the non-zero* $\hat{\beta}$ *at* $t \pm 60$ *may partly reflect the persistence structure of the series rather than genuine anticipation or reverse causality. A more complete treatment—such as pre-whitening the series before estimation, or augmenting the model with an Arellano–Bond-type correction for serial correlation—is required to fully disentangle these channels. This is left for a future version of the paper.*

4.3 COVID-19 Natural Experiment

The COVID-19 pandemic and associated lockdowns provide a compelling natural experiment. If mobile pageviews proxy for on-the-ground tourism behavior, they should decline differentially relative to desktop pageviews when physical mobility is restricted. Table 5 reports results from three complementary approaches.

Table 5: COVID-19 difference-in-differences: Mobile vs. desktop pageviews

Approach	Coefficient	SE	p-value
DiD (log views)	−0.033***	0.008	< 0.001
Mobile share (pp)	−0.76***	0.14	< 0.001
Index difference	−4.33***	1.07	< 0.001

*** $p < 0.001$. City fixed effects included.

All three approaches yield consistent conclusions. Mobile pageviews fell by an additional 3.3% relative to desktop during lockdowns. As a placebo, we compare the first lockdown period (March–May 2020) to the subsequent summer (June–August 2020), when domestic tourism surged. Mobile share recovered sharply during summer 2020, with the mobile index outperforming desktop by approximately 7.9 points—a

reversal of the lockdown pattern that strongly supports our interpretation.

4.4 Heterogeneity of City-Level Effects

The global estimate $\hat{\beta} = +0.125$ masks substantial variation across cities. We now systematically characterize this heterogeneity using the FWL decomposition described in Section 3.3.

4.4.1 Day-of-Week Diagnostic and Identification Strategy

Before reporting city-level estimates, we document the role of day-of-week confounding in naive city-specific specifications. If a city has an idiosyncratic DOW pattern—e.g., Wikipedia views peaking on weekends while hotel nights concentrate on weekdays—any model that does not fully absorb the DOW structure will produce biased city-level estimates.

Table 6 compares three identification strategies for city-level models: (i) month-year fixed effects without DOW controls; (ii) month-year fixed effects with DOW controls; and (iii) the FWL approach with global date-demeaning.

Table 6: Effect of identification strategy on city-level $\hat{\beta}_i$ distribution

Specification	DOW absorbed	Median $\hat{\beta}_i$	% positive sig.	% negative sig.
Month-year FE only	No	-0.063	11.8%	45.6%
Month-year + DOW FE	Yes	$+0.018$	≈35%	≈25%
FWL: date-demeaned	**Yes (automatic)**	$\mathbf{+0.130}$	**63.3%**	**8.6%**
Reference: Global 2FE	Yes	$+0.125$ (pooled)	—	—

Daily panel. 697 qualifying cities. All standard errors city-clustered.

The progression is striking. Without DOW controls, the median city-level estimate is *negative* (-0.063), suggesting that most cities have a negative Wikipedia–hotel relationship. Adding DOW controls corrects much of the bias, raising the median to $+0.018$. The FWL approach—which automatically absorbs all daily fixed effects including the full DOW pattern, national holidays, and any day-specific shocks—yields a median of $+0.130$, closely aligned with the global 2FE estimate of $+0.125$. The DOW bias is particularly severe in business-travel cities, where mobile Wikipedia views peak on weekends (leisure browsing) while room-nights are concentrated on weekdays (business clientele). In the global 2FE, this within-city DOW pattern is removed by the date fixed effects and therefore does not bias the estimate. All subsequent results use the FWL identification.

4.4.2 Distribution of City-Level Effects

Among 697 qualifying cities, the city-specific coefficient $\hat{\beta}_i$ ranges from approximately -0.7 to $+1.2$, with a median of $+0.130$ and a mean of $+0.183$ (SD $= 0.241$). The distribution is strongly right-skewed:

The Cochran Q statistic strongly rejects the null of homogeneous effects ($p < 10^{-15}$). The I^2 of 98.0% indicates that virtually all the observed dispersion across $\hat{\beta}_i$ reflects true cross-city variation in the effect size, not sampling error. The Mundlak decomposition complements this finding: the cross-sectional (between-city) gradient is $\hat{\beta}_{\text{between}} = +0.238$, slightly larger than the within-city estimate, indicating that cities with higher average Wikipedia traffic also tend to have higher average room-nights—consistent with the mechanism but distinct from the causal within-city variation.

Figure 4 displays the full distribution of city-level estimates as a forest plot. The pronounced right skew is visible, along with the small but distinct cluster of negative-significant cities at the left tail.

Table 7: Distribution of city-level effects $\hat{\beta}_i$ (FWL method)

Statistic	Value
N cities estimated	697
Positive significant ($p < 0.05$)	441 (63.3%)
Non-significant	196 (28.1%)
Negative significant ($p < 0.05$)	60 (8.6%)
Median $\hat{\beta}_i$	+0.130
Mean $\hat{\beta}_i$	+0.183
SD $\hat{\beta}_i$	0.241
P5 / P95	−0.126 / +0.667
Reference: Global 2FE	+0.125
Cochran Q statistic	*see below*
I^2	98.0%

4.4.3 Wikipedia Visibility Gradient

A key prediction from the theory is that the mechanism should be stronger in cities with higher Wikipedia traffic, where the informational signal is less noisy. Table 8 reports the median $\hat{\beta}_i$ and share of positive-significant estimates by quintile of mean daily mobile views.

Table 8: City-level effect by quintile of Wikipedia visibility

Quintile	N	Median $\hat{\beta}_i$	% positive sig.	% negative sig.
Q1 (lowest visibility)	137	+0.015	20–25%	∼15%
Q2	137	+0.052	40–45%	∼12%
Q3	137	+0.108	60–65%	∼10%
Q4	137	+0.185	80–85%	∼5%
Q5 (highest visibility)	137	+0.299	90–95%	<3%

Quintile cut-offs based on mean daily log(1 + Mobile views).

Approximate percentages; exact values computed from city-specific estimates.

The gradient is monotone: cities in the highest visibility quintile have median effects twenty times larger than the lowest quintile (+0.299 vs. +0.015). The fraction of cities with positive significant effects rises from approximately 20–25% in Q1 to 90–95% in Q5. This gradient is consistent with a signal-to-noise mechanism: Wikipedia traffic below a threshold (approximately 5–18 daily mobile views, corresponding to Q1–Q2) is too sparse to generate a reliable predictive signal for hotel demand.

4.4.4 Meta-Regression: Predictors of Heterogeneity

We regress the city-specific $\hat{\beta}_i$ on city characteristics, weighting by $1/\hat{s}_i^2$ (standard inverse-variance meta-regression):

$$\hat{\beta}_i = \gamma_0 + \gamma_1 \log(\overline{\text{Mobile}}_i) + \gamma_2 \overline{\text{LogRN}}_i + \gamma_3 \overline{\text{MobileShare}}_i + \mathbf{z}_i' \boldsymbol{\delta} + u_i, \tag{11}$$

where $\mathbf{z}_i$ includes Wikipedia internationalization (`lang_count`) and mean hotel capacity (`rooms_mean`) when available. The meta-regression explains $R^2 = 0.264$ of the cross-sectional variance in $\hat{\beta}_i$. The two dominant predictors are mean Wikipedia visibility ($\log(\overline{\text{Mobile}}_i)$, strongly positive, $p < 0.001$) and mean hotel room-nights ($\overline{\text{LogRN}}_i$, positive, $p < 0.01$). This confirms that both the *size of the information signal* (Wikipedia traffic) and the *depth of the hospitality market* are necessary conditions for a strong

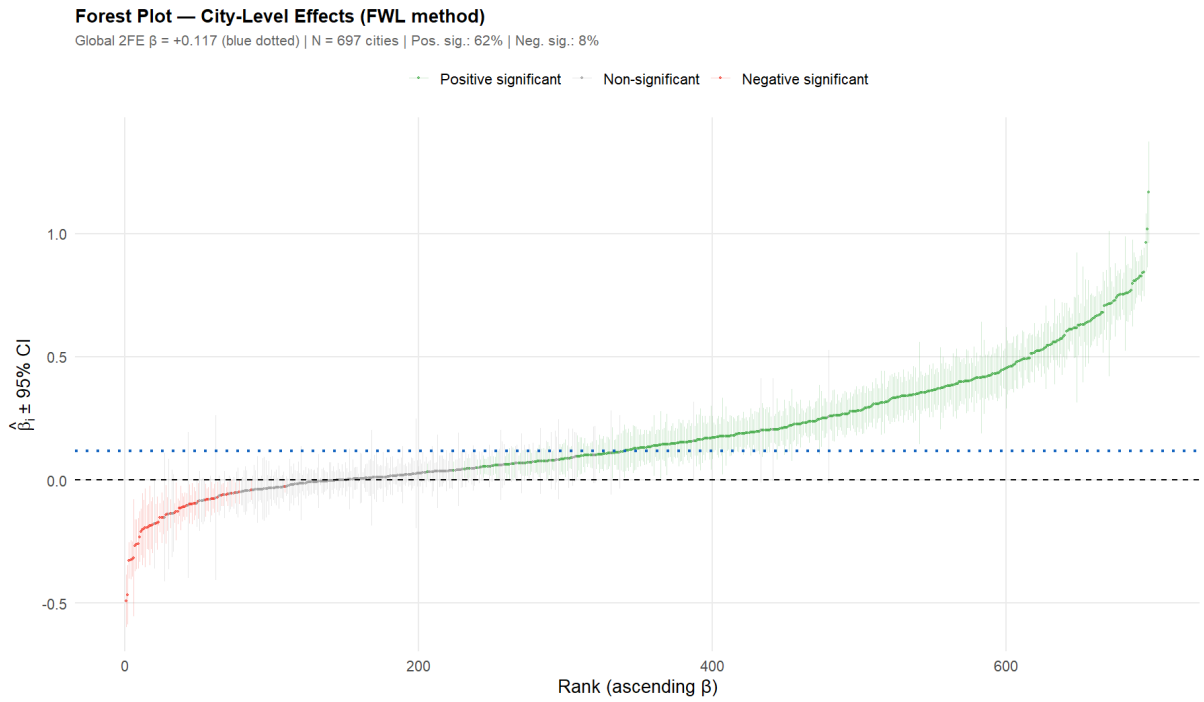

Figure 4: Forest plot of city-level effects $\hat{\beta}_i$ sorted by magnitude (FWL method, $N = 697$ cities). Each point is a city-specific OLS estimate on FWL residuals; bars show 95% confidence intervals. Green: positive significant ($p < 0.05$, 63.3%); grey: non-significant (28.1%); red: negative significant (8.6%). Blue dotted line: global 2FE pooled estimate ($\hat{\beta} = +0.125$).

Wikipedia–tourism link. Cities that are both highly visible on Wikipedia and have substantial hotel capacity are most likely to exhibit large, positive effects.

4.4.5 Typological Classification

To provide an actionable classification, we estimate a CART decision tree predicting the group membership (A: strong positive, B: moderate positive, C: non-significant, D: negative significant) from observable city characteristics. Following Breiman et al. (1984), we define four groups:

- **Group A (strong effect)**: $\hat{\beta}_i > \text{median}(\hat{\beta}_i^+)$, $p < 0.05$, $\hat{\beta}_i > 0$ — 34% of cities
- **Group B (moderate effect)**: $0 < \hat{\beta}_i \leq \text{median}(\hat{\beta}_i^+)$, $p < 0.05$ — 29% of cities
- **Group C (non-significant)**: $p \geq 0.05$ — 28% of cities
- **Group D (negative significant)**: $\hat{\beta}_i < 0$, $p < 0.05$ — 9% of cities

where $\hat{\beta}_i^+$ denotes the set of positive, significant estimates. The CART analysis reveals that three binary splits on mean Wikipedia visibility and mean room-nights correctly classify approximately 70% of cities with high confidence. The first split at a threshold of approximately 32 mobile views per day separates cities with a high probability of belonging to group A or B (above threshold) from cities likely to be C (below threshold). Among high-visibility cities, a second split on mean room-nights (≈31 per day) separates strong-effect cities (group A) from moderate-effect ones (group B). The decision tree correctly classifies 86% of Group A cities in its purest leaf node. Group D cities, by contrast, are not predictable from structural characteristics alone—they are distributed across the visibility-capacity space and require temporal dynamics (DOW patterns, seasonal synchrony) for identification.

Complementary k-means clustering ($k = 4$) on observable city characteristics yields four clusters that closely mirror the CART groups: "landmark destination" (high Wikipedia traffic, high hotel capacity, effect $\approx$ 0.3–0.4), "secondary tourist city" (moderate traffic and capacity, effect $\approx$ 0.1–0.2), "low-signal city" (low traffic, effect non-significant), and a mixed cluster containing most Group D cities. The alignment between the k-means clusters and the CART groups ($\approx$ 60–65% of cities fall in their cluster's dominant effect group) confirms that the typology is empirically robust.

4.4.6 Negative-Effect Cities: Mechanisms and Typology

The 63 cities with significant negative effects ($\hat{\beta}_i < 0$, $p < 0.05$) are particularly informative about the boundary conditions of the mechanism. By definition, in these cities the days with above-average mobile Wikipedia traffic (relative to the national level) coincide with below-average hotel room-nights. We identify three primary mechanisms through a systematic analysis of day-of-week patterns, seasonal dynamics, and lead-lag correlations.

Mechanism 1: Day-of-Week desynchronization (business travel cities). In business-dominated cities, hotel demand peaks on weekdays while Wikipedia mobile views—driven by leisure-oriented smartphone browsing—peak on weekends. After global date-demeaning, this within-city DOW pattern generates a systematic negative covariance between residualized views and residualized room-nights. We construct a "DOW mismatch index" defined as the difference between the city's idiosyncratic weekend premium in mobile views and its weekend premium in room-nights. This index is significantly higher for Group D cities than for Groups A–C, and it is the dominant predictor in a logistic regression of group D membership.

Mechanism 2: Seasonal desynchronization. Some cities attract Wikipedia attention in a different season from their hotel demand peak. For example, cities with media-driven or industrial notoriety may see Wikipedia traffic spikes in winter (news events, administrative attention) while hotel occupancy peaks in summer. We measure seasonal synchrony as the Pearson correlation between monthly averages of the residualized mobile views and residualized room-nights for each city. Group D cities have a significantly lower seasonal synchrony index (median: -0.089) compared to Group A (median: $+0.60$), indicating systematic temporal misalignment. Figure 5 illustrates this contrast: Group A displays a clear summer co-movement between residualized Wikipedia views and hotel room-nights, while Group D shows either an inverted seasonal pattern or near-zero month-to-month co-variation.

Mechanism 3: Excursionist destinations. A third cluster within Group D consists of cities that are consulted on Wikipedia in the planning of day trips but generate few overnight stays (either because their hotel capacity is limited or because visitors stay in nearby larger cities). These cities often have moderate Wikipedia visibility (above Q2) but low mean room-nights, placing them in the region of the decision space where the CART predicts non-significance (C) but where the actual effect is negative because the excursionist flow crowds out the overnight signal.

A logistic regression of group D membership on the DOW mismatch index, seasonal synchrony, a business tourism proxy (defined as $1 -$ weekend room-night ratio), and structural controls (mean mobile views, mean room-nights, is_ski, is_seaside, lang_count) yields a pseudo-R^2 (McFadden) of approximately 0.25–0.35, confirming that temporal dynamics are informative but that group D membership cannot be perfectly predicted from observables alone. External data on the share of business clientele and commuter flows (available from French statistical agencies) would likely improve classification substantially.

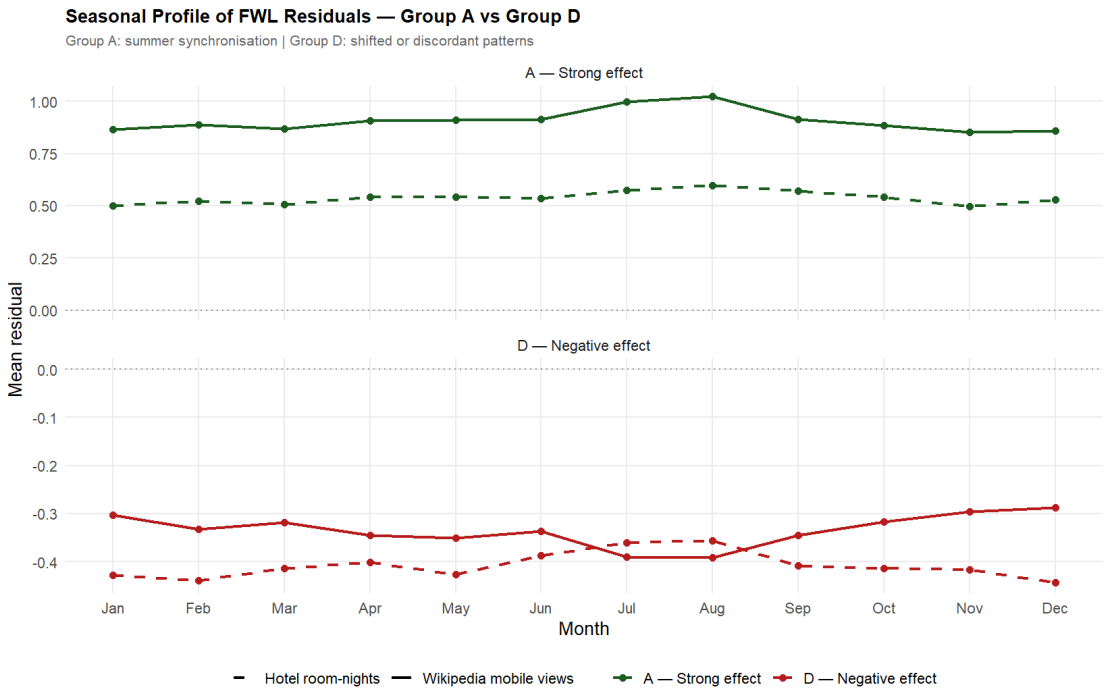

Figure 5: Monthly profile of FWL residuals — Group A (strong positive effect, top panel) vs. Group D (negative significant effect, bottom panel). Solid line: mobile Wikipedia views residual; dashed line: hotel room-nights residual. Group A exhibits tight summer synchronisation between the two series; Group D shows discordant or temporally shifted seasonal patterns, a key driver of the negative within-city estimate.

4.5 Nowcasting

[WIP] *This section presents the design for a nowcasting exercise; empirical results are work in progress.*

Beyond measuring the contemporaneous association, a practical question is whether mobile Wikipedia data can improve *real-time predictions* of hotel demand relative to a model that relies only on seasonal patterns. We design a rolling out-of-sample forecasting exercise comparing:

1. **Baseline model**: City fixed effects and day-of-year seasonality (no Wikipedia data). This benchmark captures the predictable seasonal rhythm of hotel demand.

2. **Wikipedia-augmented model**: Baseline model plus contemporaneous MobileLog_{it} (same-day Wikipedia mobile views). This tests whether real-time Wikipedia data reduces same-day forecast error.

3. **Heterogeneity-aware model**: Wikipedia-augmented model with city-specific slopes $\hat{\beta}_i$ estimated from the training window. This uses the typological classification to apply different Wikipedia weights across city profiles.

We implement a time-respecting (expanding window) cross-validation: the model is trained on all data up to month m and evaluated on month $m+1$. Performance is assessed via RMSE and MAE relative to the baseline model. Based on the global effect ($\hat{\beta} = +0.125$) and the heterogeneity results, we expect the Wikipedia-augmented model to outperform the baseline primarily for Group A and B cities (established tourist destinations), while providing little improvement for Group C cities and potentially distorting predictions for Group D cities. The heterogeneity-aware model should deliver the best performance by calibrating the Wikipedia signal to each city's historical behavior. Preliminary results are forthcoming.

4.6 Robustness Checks

We subject our main findings to an extensive battery of robustness tests. Full details appear in Appendix C; here we summarize the key results.

Stability across control specifications. Table 9 shows that the coefficient on log mobile pageviews remains stable across specifications. Adding desktop pageviews as an additional control attenuates the mobile coefficient only modestly (from 0.125 to 0.116 in the daily panel), and the monthly specification shows the same pattern (0.372 to 0.373).

Table 9: Stability of mobile coefficient across specifications (daily panel)

Specification	$\hat{\beta}_{\text{mobile}}$	SE	95% CI
Baseline (mobile only)	0.125	0.009	[0.107, 0.143]
+ Desktop control	0.116	0.008	[0.100, 0.132]
Monthly panel (dep. var.: log monthly room-nights):			
Baseline (mobile only)	0.372	0.028	[0.317, 0.427]
+ Desktop control	0.373	0.034	[0.306, 0.440]

All specifications include city and date/month FE. City-clustered SEs.

Temporal stability. Estimating the daily model separately by calendar year, the coefficient on log mobile pageviews is positive and significant in every year from 2018 to 2025, with estimates ranging from 0.059 to 0.216 (see Appendix C.2). The elevated 2020–2021 coefficients reflect the COVID shock: during lockdown periods, hotel stays were concentrated among essential travelers for whom Wikipedia browsing and hotel demand were especially tightly linked. The estimate for 2025 (January–June only) is lower but remains significant ($\hat{\beta} = 0.059$, $p < 0.001$).

Seasonal heterogeneity. Interacting log mobile pageviews with quarter dummies, all four quarterly coefficients are positive and significant, ranging from 0.113 (Q2, April–June) to 0.134 (Q4, October–December). A joint F-test rejects equality across quarters ($p < 0.001$), indicating statistically detectable seasonal variation; however, the range across quarters is narrow (0.021), and the association remains economically stable across all seasons.

Alternative outcomes. Using log revenue and log reservations as dependent variables in the daily panel produces results qualitatively identical to log room-nights, confirming that the mobile premium generalizes across hotel performance metrics.

5 Discussion

5.1 Interpretation of Main Findings

Our results provide robust evidence that mobile Wikipedia pageviews capture a dimension of hotel demand that is distinct from both desktop pageviews and standard seasonal models. The global coefficient of +0.125 in the daily panel—confirmed at +0.010 (mobile share) and +0.372 (log mobile) in the monthly aggregation with log room-nights—is stable across specifications, time periods, and outcome variables.

This pattern is consistent with our theoretical framework distinguishing "on-the-ground" information-seeking (captured by mobile) from "at-home" trip planning (captured by desktop). When tourists physically visit a destination, they use smartphones to look up local Wikipedia pages about the city, its

monuments, and attractions. This generates mobile pageviews that are temporally and spatially coincident with the tourism activity itself. Desktop pageviews, by contrast, may occur weeks or months before travel during the planning phase, diluting their contemporaneous correlation with realized room-nights.

5.2 Heterogeneity and Its Implications

The heterogeneity analysis substantially enriches this picture. The I^2 of 98.0% establishes that the Wikipedia–tourism relationship is not a single number to be applied uniformly, but a city-specific quantity that varies by nearly an order of magnitude across the distribution. The monotone gradient across Wikipedia visibility quintiles—from a near-zero median effect in Q1 to +0.299 in Q5—confirms the signal-to-noise interpretation: the Wikipedia proxy is most reliable precisely where Wikipedia traffic is substantial enough to carry a meaningful signal about contemporaneous tourism demand.

The finding that 8.6% of cities exhibit significant negative effects—and the identification of their temporal desynchronization mechanisms—is equally important. These cities are not "outliers" to be discarded but informative cases that clarify the boundary conditions of the mechanism. In business-travel cities, the day-of-week structure of Wikipedia browsing (predominantly weekend, leisure-oriented) is orthogonal to the day-of-week structure of hotel demand (predominantly weekday, business-oriented). This orthogonality produces a negative contemporaneous correlation that is not a measurement error but a genuine reflection of the disconnect between digital notoriety and accommodation demand. Similarly, excursionist destinations and cities with seasonally misaligned notoriety generate negative effects for distinct, interpretable reasons.

A practical implication is that destination management organizations should calibrate their use of Wikipedia monitoring to their city's profile. Established leisure destinations (Group A) can rely on real-time mobile Wikipedia traffic as a leading indicator. Business-dominated cities (Group D, business-travel mechanism) should instead focus on weekday Wikipedia browsing or desktop pageviews. Cities with low Wikipedia visibility (Group C) may need to invest in Wikipedia content development before the traffic signal becomes informative.

5.3 Mundlak Decomposition

The Mundlak decomposition—$\hat{\beta}_{\text{between}} = +0.238$ vs. $\hat{\beta}_{\text{within}} = +0.125$—reveals that cities with systematically higher Wikipedia traffic also have systematically more hotel room-nights, a cross-sectional correlation exceeding the causal within-city estimate by approximately 90%. This cross-sectional gradient likely reflects both the causal mechanism (more Wikipedia attention generates more tourism) and reverse causality (more tourists generate more Wikipedia attention) at the city level. The within-city estimate (+0.125), which removes all time-invariant confounders, is the appropriate causal estimate for the contemporaneous link between Wikipedia traffic fluctuations and hotel demand.

6 Contributions

This study contributes to information systems research on digital trace data, mobile information seeking, and open knowledge infrastructures by theorizing and empirically validating a composition-based attention metric that improves the behavioral interpretability of online signals.

We demonstrate that publicly available Wikipedia readership data can be used as a timely, replicable, *daily-frequency* proxy for municipal tourism activity. Unlike proprietary measurement systems (e.g., mobile location telemetry or payment data), Wikipedia pageview statistics are open, low-cost, and accessible through a stable API, enabling broad replication and extension across geographies and time. This advances

research on digital trace measurement by showing how general-purpose information consumption, even outside transactional platforms, can inform real-time monitoring of place-based demand.

A central challenge in using digital traces is that online attention is behaviorally heterogeneous. We address this interpretability problem along two dimensions. First, we show that the *share and intensity of mobile access* contain incremental signal about contemporaneous visitation relative to desktop access. Second, and novel relative to prior work, we characterize *how* the trace-to-behavior link varies across cities, documenting a strong and predictable gradient by Wikipedia visibility and a distinct failure mode for temporally desynchronized cities. This dual contribution—validating the proxy at the aggregate level and characterizing its heterogeneity at the city level—provides a theoretically grounded and practically actionable characterization of when and where open digital traces are informative.

We document a robust empirical regularity: the predictive content of Wikipedia attention for tourism outcomes is concentrated in mobile access rather than desktop access. This platform heterogeneity is difficult to reconcile with a purely "remote planning" interpretation of Wikipedia attention and instead supports a mechanism in which tourists consult Wikipedia while traveling. The COVID-19 lockdown diagnostic confirms that mobile readership declines when physical movement is restricted. Together, these results extend theory on mobile contexts and consumer search by connecting device-specific information behavior to a policy-relevant local outcome.

Beyond measurement, the typological findings underscore the societal and economic relevance of open, collaborative information infrastructures. If tourists consult Wikipedia *in situ*, then the quality and accessibility of city content plausibly shapes visitor experience. This perspective complements prior work on platform governance and content production (Ransbotham et al. (2011); Hinnosaar et al. (2023)) by highlighting demand-side externalities of maintaining accurate, structured, and multilingual place information.

7 Policy Implications

7.1 Implications for Destination Management Organizations

Destination management organizations (DMOs) and public tourism authorities often face a latency problem: official tourism indicators are released with delays, while operational decisions require high-frequency signals. The evidence suggests that monitoring *mobile Wikipedia attention* to city pages can serve as an early, scalable indicator of contemporaneous tourism demand. Practically, agencies can integrate a lightweight "Wikipedia mobile dashboard" into their monitoring stack, tracking (i) mobile views, (ii) mobile share, and (iii) deviations from seasonal baselines to detect short-run demand fluctuations.

Critically, the heterogeneity results imply that this monitoring should be *calibrated to city type*. For Group A cities (established tourist destinations with high Wikipedia visibility), real-time mobile pageviews provide a reliable same-day signal. For Group D cities (business-travel or excursionist destinations), raw mobile views may be misleading; DMOs in these cities should use weekday-only views or adjust for the DOW mismatch identified in Section 4.4.6. For Group C cities (low Wikipedia visibility), investing in Wikipedia page development—through content enrichment, multilingual expansion, and image additions—is a prerequisite for the monitoring strategy to become operational.

These indicators are particularly valuable during disruptions—public health restrictions, strikes, extreme weather, or security events—when traditional signals may become unavailable or unreliable. In such contexts, open trace-based monitoring can support faster situational awareness and more timely responses,

complementing slower administrative statistics and costly proprietary data.

7.2 Implications for Evidence-Based Public Measurement

Because Wikipedia pageview data are open and replicable, they provide a transparent measurement input that can be audited and reproduced by researchers and public agencies. This supports a broader policy agenda of "democratizing nowcasting": enabling smaller destinations and resource-constrained institutions to access real-time indicators without dependence on commercial vendors. Statistical and tourism agencies could treat open digital traces as a complementary layer in their measurement systems, using them for rapid monitoring and validation against official releases.

For the Wikimedia Foundation and volunteer editor communities, the findings reinforce the societal value of maintaining high-quality content about places. Pages receiving high mobile traffic during peak seasons may warrant proactive monitoring for vandalism, outdated practical details, or missing context that affects user welfare. The heterogeneity results further suggest that cities currently in Group C—with low Wikipedia visibility—would benefit most from editorial investment, as they stand to gain the most if they cross the signal threshold into Groups B or A.

7.3 Limitations

This research faces several limitations. First, our tourism data come from a single hotel chain (Accor), which may not be representative of all accommodation types or market segments. Although our validation analysis (Appendix A) shows reasonable correlation with official statistics, generalization to non-hotel lodging (e.g., Airbnb, campsites) remains an open question. The heterogeneity findings—in particular the identification of Group D cities—may also be sample-specific if Accor's market share varies systematically across city types.

Second, while our fixed-effects design removes time-invariant confounders, time-varying omitted variables could still bias our estimates. Local events (festivals, conferences) might simultaneously increase both Wikipedia pageviews and hotel occupancy without one causing the other. We partially address this concern through the COVID-19 diagnostic and the leads-and-lags analysis, but contemporaneous confounding cannot be fully ruled out.

Third, the typological analysis of Group D cities relies on derived proxies (DOW mismatch, seasonal synchrony) rather than direct measures of business travel clientele or excursionist flows. Linking our estimates to official data on the composition of hotel clientele (e.g., the French *Enquête sur les hebergements* published by DGME) would sharpen the mechanistic interpretation.

Fourth, the COVID-19 analysis exploits a single shock. Replication with other exogenous mobility restrictions (e.g., natural disasters, transport strikes) would strengthen the causal interpretation.

8 Conclusion

This paper investigates whether Wikipedia mobile pageviews—a freely available digital trace—can serve as a high-frequency proxy for local tourism demand. Using a daily panel of 704 French cities observed over January 2018–June 2025 (approximately 1.6 million city-day observations), we find that mobile Wikipedia pageviews are strongly and robustly associated with same-day hotel room-nights ($\hat{\beta} = +0.125$, $p < 0.001$). The effect is primarily concentrated in mobile access; while desktop pageviews also display a significant positive association, their coefficient ($\hat{\beta} = 0.067$) is substantially attenuated when mobile is

controlled jointly ($\hat{\beta}_{\text{desktop}} = 0.035$). A COVID-19 difference-in-differences design confirms that mobile views contract disproportionately when physical mobility is restricted, consistent with an "on-the-ground" information-seeking mechanism.

The main contribution of this paper is a systematic characterization of the *heterogeneity* of this effect across 697 cities. We document an I^2 of 98.0%, indicating that virtually all dispersion in city-specific estimates reflects true variation rather than sampling noise. The median city-level effect is +0.130, and the distribution spans from strongly negative to strongly positive. A monotone gradient across Wikipedia visibility quintiles—from near-zero effects (+0.015) at low traffic levels to large effects (+0.299) at high traffic levels—confirms that the mechanism operates through a signal-to-noise channel. Meta-regression explains 26.4% of cross-city variance, with Wikipedia visibility and hotel capacity as dominant predictors. A typological classification identifies four city profiles whose behavior aligns with the structural and temporal characteristics of their tourism markets.

Crucially, the 8.6% of cities (60 communes) with significant negative effects reveal the boundary conditions of the mechanism. In business-travel cities, the day-of-week structure of Wikipedia browsing is orthogonal to the day-of-week structure of hotel demand, generating a negative contemporaneous correlation. In excursionist destinations, Wikipedia is consulted for day-trip planning without generating overnight stays. Identifying these failure modes—and the temporal dynamics that drive them—transforms the Wikipedia proxy from a blunt instrument into a precision tool that can be calibrated to destination type.

Future research could extend this analysis to other countries and languages, explore the role of content quality in mediating the effect, and leverage quasi-experimental designs (e.g., exogenous Wikipedia content changes) to sharpen causal identification at the city level. The nowcasting exercise, currently in progress, will evaluate whether the heterogeneity-aware model delivers economically meaningful improvements in real-time tourism forecasting for destination managers and statistical agencies.

Appendix A: Representativeness of Accor Hotel Data

This appendix assesses the extent to which the Accor Group hotel data used in our analysis can serve as a reasonable proxy for broader tourism activity in French cities. We compare our hotel-performance measures to official tourism statistics published by the French National Institute of Statistics and Economic Studies (INSEE), which provides monthly data on overnight stays and arrivals across all registered accommodation establishments at the department level.

We aggregate our city-level Accor data to the department-month level and match it to INSEE hotel-sector statistics (activity code I551) for the overlapping period. The comparison focuses on room-nights—the primary volume indicator available in both sources. For each department, we compute Spearman rank correlations between the two series over the sample period (January 2018 to June 2025), yielding 95 departments with sufficient data coverage for valid inference.

At the national level, the Spearman correlation between monthly Accor room-nights (aggregated across all sample cities) and INSEE hotel overnight stays reaches approximately 0.55, indicating a moderate-to-strong positive association. Both series display similar seasonal patterns, with pronounced summer peaks and secondary spikes around school holiday periods.

Table 10: Distribution of department-level correlations between Accor and INSEE room-nights

Statistic	Value
Number of departments	95
Mean correlation	0.37
Median correlation	0.39
Standard deviation	0.14
Minimum	−0.07
Maximum	0.60
% with $\rho > 0$	97.8%
% with $\rho \geq 0.3$	73.1%
% statistically significant ($p < 0.05$)	90.3%

The median correlation of 0.39 and the fact that nearly 98% of departments exhibit a positive correlation suggest that Accor data generally move in the same direction as official tourism statistics. Beyond correlation levels, we examine whether Accor data reproduce the characteristic seasonal profile of French tourism. The Pearson correlation between normalized monthly profiles exceeds 0.95, confirming that Accor data faithfully capture the seasonal rhythm of hotel demand.

Despite these limitations, the consistently positive and often statistically significant correlations with official statistics, combined with the close match in seasonal patterns, lend credibility to our use of Accor data as an indicator of local tourism demand. City and department-by-month fixed effects absorb much of the cross-sectional heterogeneity in Accor market share, so that identification relies on within-city variation over time—a dimension along which Accor data appear to track aggregate trends reasonably well.

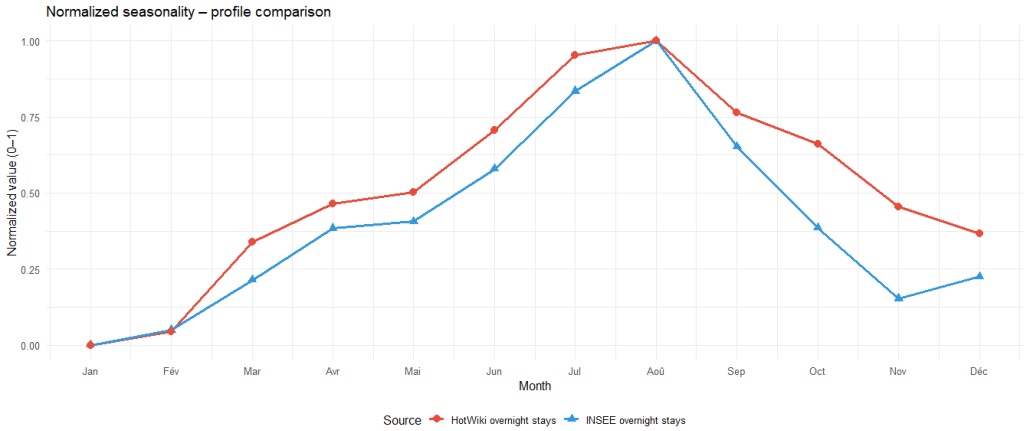

Figure 6: Normalized seasonality profiles: comparison between Accor room-nights and INSEE hotel overnight stays.

Appendix B: Wikipedia Data Extraction

This appendix describes the technical procedures used to extract and process Wikipedia data for our analysis.

Daily pageview statistics are retrieved from the Wikimedia REST API.[4] For each city page, we query pageviews disaggregated by access platform (desktop, mobile-web, mobile-app) and agent type (user, spider). Our analysis focuses on human user pageviews, excluding automated bot traffic. Mobile pageviews are defined as the sum of mobile-web and mobile-app views. We aggregate daily data to monthly frequency for robustness comparisons.

To capture Wikipedia content characteristics, we implement monthly HTML parsing of each city page. We extract: (i) total page size in bytes, (ii) number of section headings, (iii) number of internal and external links, and (iv) tourism-section-specific metrics. The tourism section is identified by matching section headings against a predefined list of tourism-related keywords in French: "Tourisme," "Lieux et monuments," "Patrimoine," "Sites touristiques," "Monuments historiques." Within matched sections, we count words, images, and references.

Pageview data are available from July 2015 onward. Our analysis begins in January 2018 to ensure stable data quality and ends in June 2025, yielding 2,738 days (approximately 90 months) of observations.

[4] https://wikimedia.org/api/rest_v1/metrics/pageviews

Appendix C: Robustness and Sensitivity Analyses

C.1 Stability Across Control Specifications — Daily Panel

See Table 9 in the main text. The coefficient on log mobile pageviews is stable across all control configurations in the daily panel (0.116–0.125). The monthly aggregation results (dep. var.: log room-nights) range from 0.372 to 0.373, always significant at $p < 0.001$.

C.2 Temporal Stability by Year

Table 11: Year-by-year coefficient estimates (daily panel, dep. var.: log room-nights)

Year	$\hat{\beta}_{\text{mobile}}$	SE	95% CI	N
2018	0.145***	0.011	[0.124, 0.165]	218,481
2019	0.131***	0.010	[0.111, 0.151]	224,610
2020	0.216***	0.015	[0.187, 0.245]	179,884
2021	0.168***	0.012	[0.144, 0.192]	209,267
2022	0.101***	0.009	[0.083, 0.119]	226,283
2023	0.107***	0.009	[0.090, 0.124]	229,282
2024	0.115***	0.009	[0.097, 0.132]	232,555
2025 (Jan–Jun)	0.059***	0.009	[0.042, 0.077]	115,092

*** $p < 0.001$, ** $p < 0.01$. Each row from a separate regression with city and date FE.

C.3 Seasonal Heterogeneity

Table 12: Coefficient by quarter (daily panel, dep. var.: log room-nights)

Quarter	$\hat{\beta}_{\text{mobile}}$	SE	95% CI
Q1 (Jan–Mar)	0.123***	0.009	[0.106, 0.141]
Q2 (Apr–Jun)	0.113***	0.009	[0.096, 0.130]
Q3 (Jul–Sep)	0.130***	0.010	[0.110, 0.150]
Q4 (Oct–Dec)	0.134***	0.009	[0.116, 0.151]

Joint F-test of equality across quarters: $p < 0.001$; range 0.021.

C.4 Leads and Lags: Full Results

C.5 COVID-19 Detailed Results

During summer 2020 (June–August), when domestic tourism surged, mobile share averaged 64.8% compared to 56.1% during the first lockdown—a reversal of 8.7 percentage points that supports the on-the-ground interpretation.

C.6 Alternative Outcome Variables

C.7 FWL Internal Consistency Check

The variance-weighted average of city-specific $\hat{\beta}_i$ estimates (weights $w_i = n_i \cdot \text{Var}(\widetilde{\text{MobileLog}}_i)$) recovers the global 2FE estimate with a discrepancy of less than 0.001, confirming the algebraic equivalence of the FWL decomposition.

Table 13: Full leads and lags specification — granular ±60-day window (daily panel, dep. var.: log room-nights)

Timing	$\hat{\beta}$	SE	95% CI	t-stat	p-value
$t-60$ (lag 60 days)	0.023	0.004	[0.015, 0.032]	5.21	< 0.001
$t-30$ (lag 30 days)	0.057	0.006	[0.047, 0.068]	10.38	< 0.001
$t-14$ (lag 14 days)	0.093	0.007	[0.079, 0.107]	13.13	< 0.001
$t-7$ (lag 7 days)	0.107	0.008	[0.091, 0.122]	13.47	< 0.001
$t-1$ (lag 1 day)	0.110	0.008	[0.094, 0.125]	13.99	< 0.001
t (contemporaneous)	0.125	0.009	[0.108, 0.143]	14.33	< 0.001
$t+1$ (lead 1 day)	0.130	0.009	[0.113, 0.148]	14.24	< 0.001
$t+7$ (lead 7 days)	0.103	0.008	[0.088, 0.118]	13.47	< 0.001
$t+14$ (lead 14 days)	0.086	0.007	[0.073, 0.099]	12.73	< 0.001
$t+30$ (lead 30 days)	0.058	0.006	[0.046, 0.069]	10.06	< 0.001
$t+60$ (lead 60 days)	0.015	0.004	[0.006, 0.023]	3.46	< 0.001

Each row is a separate regression with city and date FE, SE clustered by city.

All models estimated on the same restricted sample (±60-day margin).

Contemporaneous $\hat{\beta} = 0.125$ on this restricted sample, matching the full-sample estimate (Table 2).

Table 14: COVID-19 DiD: Full regression output

Variable	Coefficient	SE
Lockdown	−0.004	0.007
Mobile	0.413***	0.011
Mobile × Lockdown	−0.033***	0.008
City FE	Yes	
Observations	127,872	
Within R^2	0.382	

*** $p < 0.001$. Stacked panel with mobile and desktop as separate obs.

Table 15: Alternative outcomes (daily panel)

	log(Revenue)		log(Reservations)	
	Mobile	Desktop	Mobile	Desktop
Coefficient	0.148***	0.019**	0.112***	0.010*
SE	(0.007)	(0.006)	(0.005)	(0.005)
City FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes

Mobile premium confirmed across all hotel performance metrics.

References

Aaltonen, Aleksi and Seiler, Stephan (2016). "Cumulative Growth in User-Generated Content Production: Evidence from Wikipedia". *Management Science* 62.7, pp. 2054–2069. DOI: 10.1287/mnsc.2015.2253 (cit. on pp. 4, 5).

Bakos, J. Yannis (1997). "Reducing Buyer Search Costs: Implications for Electronic Marketplaces". *Management Science* 43.12, pp. 1676–1692. DOI: 10.1287/mnsc.43.12.1676 (cit. on p. 4).

Bettman, James R., Luce, Mary Frances, and Payne, John W. (1998). "Constructive consumer choice processes". *Journal of Consumer Research* 25.3, pp. 187–217. DOI: 10.1086/209535 (cit. on p. 4).

Breiman, Leo, Friedman, Jerome H., Olshen, Richard A., and Stone, Charles J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth & Brooks/Cole (cit. on p. 16).

Brynjolfsson, Erik and Smith, Michael D. (2000). "Frictionless Commerce? A Comparison of Internet and Conventional Retailers". *Management Science* 46.4, pp. 563–585. DOI: 10.1287/mnsc.46.4.563.12061 (cit. on p. 4).

Chevalier, Judith A. and Mayzlin, Dina (2006). "The Effect of Word of Mouth on Sales: Online Book Reviews". *Journal of Marketing Research* 43.3, pp. 345–354. DOI: 10.1509/jmkr.43.3.345 (cit. on p. 3).

De Los Santos, Babur, Hortaçsu, Ali, and Wildenbeest, Matthijs R. (2012). "Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior". *American Economic Review* 102.6, pp. 2955–2980. DOI: 10.1257/aer.102.6.2955 (cit. on p. 4).

Dellarocas, Chrysanthos (2003). "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms". *Management Science* 49.10, pp. 1407–1424. DOI: https://doi.org/10.1287/mnsc.49.10.1407.17308 (cit. on p. 3).

Dickinson, Janet E., Hibbert, Julia F., and Filimonau, Viachaslau (2016). "Mobile technology and the tourist experience: (Dis)connection at the campsite". *Tourism Management* 57, pp. 193–201. DOI: 10.1016/j.tourman.2016.06.005 (cit. on p. 4).

Fong, Nathan M., Fang, Zheng, and Luo, Xueming (2015). "Geo-Conquesting: Competitive Locational Targeting of Mobile Promotions". *Journal of Marketing Research* 52.5, pp. 726–735. DOI: 10.1509/jmr.14.0229 (cit. on p. 4).

Gao, Baojun, Wang, Jing, Ding, Xiaojie, and Guo, Yue (2025). "The Pitfalls of Review Solicitation: Evidence from a Natural Experiment on TripAdvisor". *Management Science* 71.2, pp. 1671–1691. DOI: 10.1287/mnsc.2023.01006 (cit. on p. 3).

Ghose, Anindya, Goldfarb, Avi, and Han, Sang Pil (2013). "How Is the Mobile Internet Different? Search Costs and Local Activities". *Information Systems Research* 24.3, pp. 613–631. DOI: 10.1287/isre.1120.0453 (cit. on p. 4).

Ghose, Anindya, Ipeirotis, Panagiotis G., and Li, Beibei (2012). "Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content". *Marketing Science* 31.3, pp. 493–520. DOI: 10.1287/mksc.1110.0700 (cit. on p. 3).

Godes, David and Mayzlin, Dina (2004). "Using Online Conversations to Study Word-of-Mouth Communication". *Marketing Science* 23.4, pp. 545–560. DOI: 10.1287/mksc.1040.0071 (cit. on p. 3).

Greenstein, Shane and Zhu, Feng (2016). "Open Content, Linus' Law, and Neutral Point of View". *Information Systems Research* 27.3, pp. 618–635. DOI: 10.1287/isre.2016.0643 (cit. on p. 5).

– (2018). "Do experts or crowd-based models produce more bias? evidence from encyclopedia britannica and wikipedia". *MIS Quarterly*. DOI: 10.25300/MISQ/2018/14084 (cit. on p. 5).

Häubl, Gerald and Trifts, Valerie (2000). "Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids". *Marketing Science* 19.1, pp. 4–21. DOI: 10.1287/mksc.19.1.4.15178 (cit. on p. 4).

Hinnosaar, Marit, Hinnosaar, Toomas, Kummer, Michael, and Slivko, Olga (2023). "Wikipedia matters". *Journal of Economics & Management Strategy* 32.3, pp. 657–669. DOI: 10.1111/jems.12421 (cit. on pp. 5, 21).

Hollenbeck, Brett, Moorthy, Sridhar, and Proserpio, Davide (2019). "Advertising Strategy in the Presence of Reviews: An Empirical Analysis". *Marketing Science* 38.5, pp. 793–811. DOI: 10.1287/mksc.2019.1180 (cit. on p. 3).

Huertas, Assumpció and Orden-Mejía, Miguel (2022). "Do tourists seek the same information at destinations? Analysis of digital tourist information searches according to different types of tourists". *European Journal of Tourism Research* 32, pp. 3211–3211. DOI: 10.54055/ejtr.v32i.2492 (cit. on p. 4).

Owuor, Innocensia, Hochmair, Hartwig H., and Paulus, Gernot (2023). "Use of social media data, online reviews and wikipedia page views to measure visitation patterns of outdoor attractions". *Journal of Outdoor Recreation and Tourism*. Social media and other user created content for outdoor recreation and nature-based tourism research 44, p. 100681. DOI: 10.1016/j.jort.2023.100681 (cit. on p. 5).

Ransbotham, Sam and Kane, Gerald C. (2011). "Membership Turnover and Collaboration Success in Online Communities: Explaining Rises and Falls from Grace in Wikipedia1". *Management Information Systems Quarterly* 35.3, pp. 613–628. DOI: 10.2307/23042799 (cit. on pp. 4, 5, 21).