ParagraphS
Paragraphs Environmental Design

Review Article

DOI: 10.5281/zenodo.24157888

OPEN ACCESS

***CORRESPONDENCE**
Emirhadi Suganda
emirhadisug42@gmail.com

RECEIVED: June 2024
ACCEPTED: August 2024
PUBLISHED: September 2024

ISSN-L 2960-2874

VOLUME 2, ISSUE 2
2024

CITATION
Suganda E. (2024). Construction Period and Office Building Energy Use: A Comparative Analysis of Benchmarking Records. Paragraphs Environmental Design, 2(2), 26-41. DOI: 10.5281/zenodo.24157888.

COPYRIGHT
© 2024 Suganda E. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Construction Period and Office Building Energy Use: A Comparative Analysis of Benchmarking Records

Emirhadi Suganda*

Professor Emeritus, Department of Architecture,
Universitas Indonesia

Abstract

Construction period is frequently treated as a practical indicator of office building energy performance because it records the historical conditions under which envelopes, services, and regulatory requirements were established. Yet measured benchmarking records do not consistently rank newer offices above older ones, and apparently conflicting findings often concern different fuels, populations, and analytical questions. This article develops a critical integrative comparison of research and public benchmarking documentation published before 1 May 2023. It examines office evidence from New York City, England and Wales, and Seoul, alongside United States national survey and ENERGY STAR documentation. The comparison distinguishes chronological age from construction cohort, operational demand from technical efficiency, and observed associations from intervention effects. Published studies reveal older-stock advantages in some settings, divergent electricity and thermal trends in others, and weak construction-year relationships elsewhere. These patterns cannot be combined into a universal age coefficient. The article proposes a framework that separates stock energy burden, conditional peer performance, and causal evaluation, with explicit attention to metering boundaries, floor-area definitions, survey weights, refurbishment, occupancy, and reporting selection. Construction period emerges as a useful historical descriptor and diagnostic starting point whose interpretation depends on the outcome and comparison population. Its strongest policy use is to organize investigation and identify heterogeneous retrofit opportunities; it is insufficient by itself to rank individual offices, evaluate building codes, or justify replacement of existing stock.

Keywords: office buildings; construction period; energy benchmarking; operational energy; construction cohort; building performance

1. Introduction

An office building's construction date appears to offer a simple explanation for its energy performance. More recent construction can incorporate improved insulation, efficient equipment, and better control systems, making lower measured consumption a reasonable expectation. However, the same historical transition can involve larger glazed areas, deeper floor plates, more mechanical services, and more intensive occupation. Consequently, construction period combines changes in technical capability with changes in the services buildings provide. This article argues that its explanatory value depends on disentangling these components rather than assuming that a calendar date represents a single technological improvement. The resulting question is precise: under what conditions do benchmarking records support a meaningful relationship between construction period and office energy use, and what conclusions remain unsupported when those conditions are absent?

Measured benchmarking records provide an essential basis for answering that question because they describe energy used during operation. They can reveal outcomes that design calculations or equipment specifications alone cannot establish. Nevertheless, a measured outcome is not automatically a comparable outcome: two annual energy totals can cover different occupied areas, tenant loads, weather conditions, or property boundaries. Hsu (2014) shows why self-reported benchmarking requires attention to input quality and the purposes for which the information is used. The appropriate response is to treat measurement and comparability as separate achievements. A reliable meter establishes consumption within its boundary; a defensible comparison additionally establishes that the selected buildings and denominators answer the same research question.

The empirical literature already challenges a universal progression from old inefficiency to new efficiency. Kontokosta (2015) reports an association between more recent construction and higher source energy intensity among Manhattan offices, whereas Armitage et al. (2015) identify opposing historical trends in electricity and fossil-thermal use in public offices in England and Wales. These findings are informative, but neither supplies a timeless ranking of construction cohorts. Each reflects a particular stock, reporting period, and energy definition. Their joint implication is therefore methodological as well as substantive: disagreement across studies may expose different mechanisms or comparison boundaries rather than an error that a larger pooled average would necessarily resolve.

The distinction matters because several practical decisions are routinely compressed into a single question about whether older buildings perform worse. A municipality may need to locate the largest absolute energy demand, an owner may need to identify operational underperformance, and a regulator may need to evaluate a code change. The buildings relevant to these tasks can overlap, but the required comparisons differ. Consistent with the separation of description, prediction, and causal inference developed by Hernán et al. (2019), this article treats the purpose of analysis as prior to model selection. An answer about present stock composition should not silently become an estimate of what would happen if a particular building were renovated or replaced.

The article's contribution is an interpretive framework connecting those purposes to the practical structure of benchmarking records. Existing studies offer valuable local results, while methodological and official documents explain different aspects of normalization, sampling, and measurement. Their findings do not, by themselves, create a common basis for comparing construction periods across jurisdictions. The analysis therefore identifies the conditions under which conclusions can be carried from one setting to another and specifies a transparent analytical protocol for future record-level research. Its originality lies in that comparative synthesis and the resulting research design; it does not claim a newly estimated construction-period effect or a new dataset of office energy consumption.

2. Review scope and analytical approach

The evidence base consists of a purposive critical review of publications available before 1 May 2023. Searches combined terms concerning office buildings, construction year, building age, vintage, measured energy, and benchmarking, followed by examination of publisher records, university repositories, official reports, and cited

methodological studies. This approach prioritizes traceable evidence directly relevant to the comparison, although it cannot establish exhaustive coverage of every language, jurisdiction, or publication database. Accordingly, the article is an integrative review rather than a systematic review or meta-analysis. No comprehensive screening count or pooled effect size is reported, and the strength of each conclusion remains tied to the evidence actually available.

The central empirical sources were selected for their examination of measured office consumption and construction-related characteristics. Official documentation was included to establish how public records, energy metrics, and reference populations are constructed; broader commercial-building studies were used only where they illuminate a specific mechanism or methodological dispute. Such supporting evidence broadens interpretation, but it cannot substitute for an office-specific cohort estimate. The comparison therefore distinguishes direct office evidence from contextual evidence throughout. A study of commissioning, for example, can demonstrate that operating condition changes independently of original construction date without establishing which historical office cohort offers the greatest savings potential.

Source verification checked author names, titles, publication years, journal or institutional provenance, and persistent identifiers against primary publication or repository records. Findings were drawn from accessible full texts and official documents where available; when access was limited to an abstract or indexed sections, discussion was restricted to claims visible there. Verification makes the bibliography auditable, but it does not remove limitations in the original research. This distinction is especially important for numerical results: a correctly cited estimate can still depend on selective reporting, incomplete controls, or a narrow sample. The synthesis consequently evaluates what each estimate represents as carefully as whether the reference exists.

Comparison proceeded across population, reporting year, construction-period definition, energy boundary, statistical adjustment, and inferential purpose. These dimensions expose why results that appear inconsistent may answer different questions. They also limit aggregation: an unadjusted electricity comparison cannot be treated as the same outcome as a conditional source-energy estimate. Rather than calculate a misleading common effect, the article compares the direction and interpretation of findings within their stated boundaries. The analytical recommendations that follow are deductions from this comparison and general measurement logic, explicitly distinguished from empirical results. They offer a protocol for testing construction-period relationships without suggesting that the protocol has already been executed here.

3. Construction period as a historical and analytical variable

Construction period can usefully summarize persistent features of a building. Original structural form, orientation, depth, and façade arrangement may constrain later energy choices long after individual pieces of equipment have been replaced. Yet a period label also conceals substantial variation within the cohort, including differences in design quality, investment, climate, and intended market. Its defensible interpretation is therefore probabilistic: it can indicate the likelihood of certain characteristics, not guarantee their presence. A comparison should use construction period to organize hypotheses about relevant features and then examine those features where the records permit, rather than treating the label as a complete physical description.

Chronological age and construction cohort must also be separated conceptually, even though they are algebraically linked. In any observation year, building age equals observation year minus construction year. A cross-sectional contrast between young and old buildings is consequently a contrast between different historical cohorts, not an observation of the same building becoming older. Adding several reporting years expands the evidence but does not dissolve the identity among age, period, and cohort. Thus, a study must state which variation it interprets and which assumptions permit that interpretation. Describing a cross-sectional coefficient as physical deterioration over time would attribute a process that the comparison has not directly observed.

The distinction becomes especially consequential when contemporary technology is installed in an old shell. An office constructed decades earlier can retain its original structural constraints while receiving new heating equipment, lighting, glazing, and controls. Conversely, a recent building can operate poorly despite its newer components. Mills (2011), drawing on commissioning projects in nonresidential buildings, documents opportunities to correct operational deficiencies in both new and existing properties. The project evidence is not a representative comparison of office cohorts, but it establishes the relevance of current operating condition. Construction year should therefore be retained alongside dated refurbishment and commissioning information, rather than allowed to stand in for every component's age.

Different building services can also produce countervailing energy responses to the same historical change. Better insulation may reduce heating requirements, while more extensive cooling or ventilation increases electricity consumption. Such a mechanism is physically plausible, but the annual total alone cannot show which component changed or whether the trade-off reflects improved comfort, poor operation, or another cause. A defensible construction-period analysis should therefore begin with separate electricity and thermal-fuel outcomes wherever possible. Fuel separation still does not identify end uses: electricity can supply heating, cooling, ventilation, and office equipment. Combining fuels remains useful for accounting, but should follow examination of their components. Otherwise, a stable total could conceal substantial improvement in one end use and deterioration in another.

Energy efficiency itself requires a specified service. Low annual energy per unit floor area can indicate efficient delivery of comfort and workplace services, but it can also reflect reduced occupation, short operating hours, or inadequate conditioning. Conversely, a more intensively occupied building may consume more per square metre while providing more workplace activity from the same area. These possibilities prevent a simple identification of consumption with technical quality. The synthesis is to retain observed consumption as an important environmental outcome while using service and occupancy information to interpret performance. Neither a low meter reading nor an attractive normalized ranking should be accepted as a complete account of what the building accomplishes.

Historical cohorts may embody different service expectations as well as different equipment. Andrews and Krogmann (2009), using United States commercial-building surveys, found that adoption of several technologies did not straightforwardly correspond to lower annual energy intensity. Their study concerns commercial stock rather than a controlled office-vintage experiment, so it does not establish a causal explanation for any particular office comparison. It nevertheless cautions against equating the presence of more recent technology with achieved savings. For construction-period research, the useful hypothesis is that technical improvements and changing demand can coexist. Testing that hypothesis requires operating and service information rather than inferring a rebound effect from aggregate consumption alone.

Three analytical targets follow from these distinctions. Descriptive analysis concerns the energy burden of the offices currently represented in each cohort; conditional analysis concerns differences among buildings with specified comparable characteristics; causal analysis concerns the effect of a defined intervention. Each target can be legitimate, but none is interchangeable with the others. The same variable may be relevant to one target and inappropriate as an adjustment for another. The remainder of the article uses this separation to interpret published comparisons and to develop a design in which the reported result can be traced back to a clearly stated question.

4. Comparative evidence from office benchmarking records

Kontokosta (2015) provides a particularly relevant conditional comparison using 2010 energy data merged with property and market information for 685 Manhattan office properties. Buildings completed from 1990 onward had approximately 40% higher weather-normalized source energy intensity than the pre-1930 reference group after the included controls. The

result challenges a simple newer-is-better assumption, although the paper recognizes possible omitted tenant characteristics and its newest building was completed in 2007. It consequently cannot evaluate subsequent construction requirements or demonstrate that age itself reduces demand. The appropriate conclusion is narrower and more useful: within this historical market and specification, the newer cohort's measured consumption advantage did not materialize, making operating intensity and historical building form necessary subjects for further investigation.

The broader municipal record supports attention to the shape of the relationship. Urban Green Council's (2016) report on New York City's 2013 energy and water use presents a construction-decade comparison in which older office cohorts generally have lower weather-normalized source intensity than postwar cohorts, followed by a decline after the 1990s. The report discusses possible physical and operational explanations, but the figure is descriptive and does not isolate their contributions. This evidence therefore qualifies both linear improvement and linear deterioration narratives. A useful comparison must allow some recent cohorts to improve relative to their immediate predecessors while still consuming more than much older cohorts; one age slope would obscure that historical pattern.

Armitage et al. (2015) offer a complementary comparison using Display Energy Certificate records for approximately 2,600 public offices in England and Wales, with age information available for approximately 550. Newer offices displayed lower typical fossil-thermal consumption but higher electricity use and higher typical carbon emissions under the study's conversion factors. The smaller age-observed subset and public-sector setting restrict generalization, while historical emissions factors prevent direct transfer of the carbon ranking to other electricity systems. The study nevertheless establishes a crucial comparative point: construction-period conclusions can change with the outcome chosen. A newer cohort can show lower fossil-thermal consumption without improving on electricity or the historical carbon measure, so apparently contradictory rankings can describe the same underlying stock.

Building height adds another dimension to that interpretation. Godoy-Shimizu et al. (2018) examined 611 public and commercial offices in England and Wales and reported higher energy intensity in taller buildings; more recently constructed offices also had higher electricity intensity without a compensating reduction in fossil-fuel intensity. These associations implicate morphology and servicing in the comparison, but they do not show that height or construction date independently caused the observed consumption. The study is most useful here as evidence that date should be interpreted within a building's physical and operational configuration. Treating construction year as an isolated indicator would discard information needed to explain why offices of different periods can have different energy demands.

Ahn et al. (2017) provide a geographically distinct result from 4,625 offices in Seoul, using 2015 electricity and gas records and primary-energy indicators. Their descriptive analyses show weak relationships between construction year and energy intensity. Although the authors question the effectiveness of prescriptive requirements, the absence of a strong cross-sectional relationship does not isolate the effect of those requirements from occupancy, systems, refurbishment, or other differences. The measured pattern is therefore a weak observed vintage association, not proof that regulation had no technical benefit. Its contribution to the comparison is to show that the Manhattan relationship cannot be assumed to recur in another urban building stock with different histories and operating conditions.

Table 1 summarizes the direct office comparisons without imposing a common numerical effect. Such a matrix makes differences in coverage and outcome visible, although it necessarily compresses details that remain important in the original studies. It should therefore be read as an orientation to the evidence rather than a league table of countries or construction eras. The consistent finding across these heterogeneous records is not a single direction of age effect. It is the insufficiency of construction date, on its own, to establish a

universally ordered progression in measured office energy performance.

Table 1. Comparison of published office evidence

Source	Population	Outcome and comparison	Main interpretive limit
Kontokosta (2015)	685 Manhattan offices	Adjusted source EUI; newer cohort higher	Historical market; tenant confounding
Armitage et al. (2015)	England and Wales public offices	Electricity and fossil-thermal trends differ	Age known for only part of sample
Godoy-Shimizu et al. (2018)	611 offices in England and Wales	Energy associated with height and vintage	Correlated building characteristics
Ahn et al. (2017)	4,625 Seoul offices	Weak descriptive vintage relationship	No identified causal code effect
Urban Green Council (2016)	Covered New York City properties	Nonmonotonic office cohort pattern	Descriptive reporting population

The studies' differences are analytically productive because they suggest questions that a universal age coefficient would suppress. A source-energy association in one city, an electricity-thermal divergence in another stock, and a weak overall correlation elsewhere can all arise without logical contradiction. Nevertheless, recognizing heterogeneity should not become a reason to make every result immune to challenge. A satisfactory explanation must specify which differences in measurement, services, or stock composition could produce the observed contrasts and identify the information needed to test them. The synthesis therefore treats heterogeneity as a research problem with observable implications, rather than as an unrestricted justification for whatever a local dataset happens to show.

National data can provide context for these local results, provided that the target population is preserved. The 2018 Commercial Buildings Energy Consumption Survey table reports office major-fuels intensity of 65.6 kBtu per square foot for survey-weighted aggregate consumption divided by survey-weighted aggregate floor area, compared with its published survey median building-level intensity of 53.5 (U.S. Energy Information Administration [EIA], 2022). These are different summaries, not inconsistent measurements. Moreover, construction-period rows elsewhere in that table describe all commercial buildings, not offices crossed with construction year. The table can illustrate distributional and weighting issues, but it cannot independently supply office-specific cohort means. Producing those means would require the appropriate joint records or a published joint tabulation.

Predictive research further demonstrates the usefulness, and the limits, of combining benchmarking with property information. Kontokosta and Tull (2017) used New York City records and building attributes to model energy consumption across a much broader city stock. Their expanded coverage consists partly of predictions, however, and should not be confused with a comparably large set of direct measurements or an office-only causal estimate. The lesson for construction-period analysis is to preserve the distinction between an observed record and an inferred value. Predictions can help locate potential demand and guide data collection, while conclusions about historical performance should remain transparent about the evidence actually measured.

5. Energy metrics and the meaning of a cohort comparison

Site energy intensity is an intuitive outcome because it relates energy delivered within a specified boundary to the corresponding floor area. It can support management of the consumption recorded by building meters, but it

does not incorporate upstream energy requirements in the same way as source energy. Source intensity adds conversion factors for those requirements, which can alter rankings where cohorts have different fuel mixes. The appropriate synthesis is to publish the component fuels and state the accounting boundary before interpreting a cohort contrast. Neither metric is intrinsically sufficient for every purpose: the relevant measure depends on whether the question concerns building operation, energy-system resources, or another explicitly defined objective.

The historical LEED debate illustrates the practical importance of these choices. Newsham et al. (2009) reported lower average energy use per floor area among the certified buildings they examined, alongside substantial variation and a sizeable share that consumed more than comparison buildings. Their analysis concerns a mixed commercial and institutional sample and historical certification schemes, so it cannot be translated directly into an office construction-period effect. Its relevance is the coexistence of an average advantage with overlap between individual outcomes. In a cohort analysis, the equivalent lesson is that a favorable group mean neither guarantees the performance of a particular member nor removes the need to inspect the distribution.

Scofield (2009), examining related certification evidence, emphasized weighting and site-versus-source definitions and found that the floor-area-weighted site- and source-energy differences for the office subset were not statistically significant. This result warrants scrutiny of the estimand, but a failure to detect a difference does not establish physical equivalence or invalidate all certification strategies. Taken together, the two analyses show why a disagreement cannot be resolved by comparing headline conclusions alone. For construction-period research, the corresponding discipline is to specify whether the result concerns the average building, the typical building, or the energy burden of the floor area represented. Those quantities can legitimately produce different answers from the same records.

For a cohort, the arithmetic mean of individual energy intensities gives each building equal weight. Aggregate intensity, calculated as total energy divided by total floor area, instead weights each building's intensity by its area. Both measures are mathematically legitimate, but they answer different questions whenever size and intensity vary together. A small office and a large tower have equal influence on the first measure and very different influence on the second. The resulting synthesis is straightforward: use building-weighted summaries for questions about buildings and aggregate intensity or total energy for questions about stock demand. Report both when policy discussions move between those scales, rather than allowing the choice of average to remain implicit.

The median is valuable because a few very intensive properties can pull a mean away from the centre of the distribution. Yet the same high-consuming properties may account for a large share of total demand and therefore matter greatly for intervention planning. A median should consequently complement, rather than replace, upper-tail and total-consumption information. Cohort comparisons should show sample counts, central tendency, dispersion, and the amount of floor area represented. This combination distinguishes a widespread shift in performance from a difference driven by a small group of large or unusual buildings. It also makes clear why statistical robustness to extremes and practical attention to major consumers are compatible objectives.

The ENERGY STAR office score introduces another kind of outcome. The July 2019 technical reference describes a national percentile comparison based on a source-energy model using 2012 CBECS data and adjustments for specified building, operating, and weather characteristics; construction year is not among its listed adjustments (U.S. Environmental Protection Agency [EPA], 2019b). A score is therefore neither a directly metered intensity nor an explicit age allowance. Its omission of construction year also does not prove that construction period is unrelated to demand. In comparative research, scores can describe performance within that model's reference framework, while measured fuel use is needed to explain what physical consumption differences lie behind the ranking.

Weather normalization reduces one source of variation but must be interpreted at the scale it

actually addresses. EPA (2020) distinguishes adjustment to a property's own normal weather from adjustment for differences in climate between locations. Thus, selecting weather-normalized intensity does not automatically make an office in one climate comparable with an office in another. The synthesis is to treat weather year and climate region as separate dimensions of comparison. Within-city cohort contrasts may require relatively little cross-climate adjustment, whereas international comparisons require explicit attention to the baseline heating and cooling conditions against which buildings operate. A normalized label cannot substitute for that geographic reasoning.

Changes in a benchmarking model can also change the apparent performance history. EPA's review of its office model led to the reintroduction of a heating-degree-day adjustment in July 2019, potentially increasing some scores without a physical energy intervention (EPA, 2019a). The change improved a particular assessment framework, but it also means that score movement across versions cannot be interpreted solely as movement in building performance. A longitudinal cohort comparison should record the model version and, where feasible, apply a consistent definition retrospectively. Otherwise, an administrative revision could be mistaken for improvement concentrated in the cohorts or locations most affected by the revision.

Carbon intensity adds an environmental dimension that operational energy alone cannot fully express. The same fuel consumption can imply different emissions when electricity generation or emissions factors differ across places and reporting years. Nevertheless, changing the carbon factor does not itself demonstrate a change in the building's physical energy efficiency. The comparison should therefore keep consumption and emissions calculations visible as separate layers, identifying the factors and dates used. This permits a policy-relevant carbon assessment while preserving the ability to ask whether an apparent cohort improvement arose from building operation, fuel switching, or the energy supply against which consumption was evaluated.

6. A defensible protocol for comparative record analysis

The first requirement of a future empirical study is a declared target population. An analysis of all offices in a city, offices above a reporting threshold, and offices with complete usable reports addresses three different populations. Mandatory reporting can make coverage extensive, but a statutory requirement does not guarantee complete compliance or eliminate exclusions introduced during cleaning. The protocol should therefore identify the eligible stock, reporting stock, and final analytical stock separately. Reporting the number and floor area retained by construction cohort would show whether an apparent vintage pattern emerges partly from differential coverage. This is a proposed transparency requirement, not a finding that any particular cohort is systematically missing in the reviewed datasets.

The next requirement is to make the unit of observation consistent with the metered boundary. A record may represent a single building, several buildings sharing meters, or a property containing multiple activities. Matching it to one construction year is convenient, but potentially misleading if different structures or extensions belong to different periods. Such cases should receive a documented aggregation rule or be analysed separately, with their contribution to total energy retained in the descriptive account. The goal is to avoid assigning precise historical meaning to an energy total whose physical footprint spans several cohorts. A smaller well-defined office sample can answer a narrower question more credibly than a larger sample with ambiguous units.

Floor-area verification deserves equal attention because an intensity ratio depends on its denominator as much as its numerator. Gross area, conditioned area, rentable area, and occupied area can describe different portions of the same property. Substituting one for another may change a ranking even when the energy total is correct. A comparison should consequently document the area convention, reconcile major revisions, and avoid pairing whole-property energy with an office-only denominator unless a valid allocation is available. This requirement does not imply that one area definition is universally superior. It

means that the selected definition must match the intended service boundary and remain consistent across the cohorts being compared.

The office classification should then be examined for uses that materially alter consumption. A property whose principal activity is office work may also contain a data centre, catering facilities, retail space, or unusually intensive equipment. Excluding every complex property would improve homogeneity but could also remove a substantial and historically uneven part of the office stock. The proposed solution is to distinguish a clearly defined office comparison from supplementary mixed-use analyses and report the consequences of that distinction. Where disaggregation is unavailable, the limitation belongs in the estimand itself. A comparison of office-dominant properties should not be described as a measurement of office workspace consumption alone.

Construction periods should reflect a defensible historical rationale rather than convenient equal-width bins alone. Locally relevant changes in building form, servicing, or regulation may justify cohort boundaries, while broad decades can remain useful for descriptive communication. Yet selecting boundaries after inspecting which split produces the strongest result risks manufacturing an apparently clear historical transition. The protocol should therefore specify its primary grouping before estimation and evaluate alternative groupings as sensitivity checks. It should also show continuous construction-year patterns where sample density permits. Agreement across reasonable definitions strengthens a descriptive conclusion; disagreement identifies dependence on categorization rather than an excuse to report only the most favourable specification.

Refurbishment information requires a similarly explicit treatment. A single renovation flag can capture anything from a limited alteration to a comprehensive replacement of major systems, making it an unreliable measure of comparable technical renewal. Omitting refurbishment entirely, however, risks attributing present equipment performance to original construction. A useful record structure would retain original completion year and separately record the dates, scope, and affected systems of substantial interventions. Where only an uncertain flag is available, results should distinguish observed refurbishment categories without equating them with common retrofit depth. The comparison can then acknowledge partial information instead of replacing a complicated physical history with an unjustifiably precise adjusted age.

Data-quality rules should be tied to identifiable errors and plausible physical boundaries. Impossible dates, inconsistent units, duplicate submissions, and partial-year records warrant investigation, but unusually high consumption may be a real characteristic of the stock. Automatic deletion of extreme intensities can therefore suppress precisely the properties that matter for energy policy. The protocol should report exclusions by reason and cohort, distinguish corrected errors from retained extremes, and show sensitivity to reasonable outlier treatments. The central requirement is traceability: another researcher should be able to determine whether a reported cohort difference survives an alternative defensible cleaning rule, rather than having to accept the cleaned dataset as an unexplained starting point.

Missing data should be evaluated in relation to the source's design. The CBECS user guide documents sampling and replicate weights, imputation flags, and the use of supplied imputed values in national estimation (EIA, 2023). These features make a simple complete-case analysis an inadequate default for that survey, although municipal disclosure data may require a different approach. Accordingly, the protocol should follow the official design for survey estimates and separately examine missingness in administrative records. Occupancy missingness can be assessed within known cohorts; missing construction year must instead be examined against other observed characteristics, since the cohort is then generally unknown. More elaborate imputation does not remove the need to explain the assumptions under which missing records are represented.

Descriptive estimation should precede adjustment because the raw distribution defines the phenomenon requiring explanation. Cohort counts, floor area, total consumption, fuel-specific intensities, and dispersion establish whether the issue is widespread demand, a few large consumers, or a difference between electricity and

thermal use. These summaries cannot identify causal mechanisms, but they prevent a regression coefficient from becoming detached from the observed stock. The protocol should therefore present unadjusted results as substantive findings in their own right, alongside any later conditional comparison. Their interpretation remains the actual reported population under actual operation, rather than a hypothetical population in which operating conditions have been standardized.

Conditional modelling can then ask whether cohort differences remain among offices with specified comparable characteristics. Flexible size terms, location, weather, operating hours, occupancy, and relevant use categories are reasonable candidates where measured adequately. Nevertheless, including more variables does not necessarily make the construction-period coefficient more meaningful. Current servicing and morphology may be part of the historical pathway represented by construction period, while some operating variables may be measured with substantial error. The protocol should use a sequence of models with explicit purposes: an initial stock description, a service-adjusted comparison, and a further physical-characteristic comparison. A change across them shows sensitivity to conditioning, not automatic proof of mediation or confounding removal.

The choice of statistical method should follow those purposes. Chung (2011) reviews benchmarking approaches with different properties, while Hsu (2015) shows the usefulness of regularization and interactions in predicting building consumption. These contributions support considering alternatives to a single rigid regression, although predictive success does not convert a model into a causal design. For the proposed comparison, regression, quantile methods, or flexible prediction can each be justified if they clarify a stated target and are evaluated accordingly. A method should not be selected solely because it gives the clearest age coefficient. The study must explain what the fitted quantity means and how uncertainty or model dependence affects its interpretation.

Comparable support is essential when interpreting an adjusted difference. If the oldest cohort consists mainly of small, naturally ventilated buildings and the newest mainly of large towers, a model may predict counterfactual combinations that are scarcely represented in the records. More flexible fitting does not supply observations in those missing combinations. The protocol should therefore examine overlap in major characteristics, identify where comparisons require extrapolation, and provide restricted analyses where justified. Such restrictions change the target population and must be disclosed. A conclusion about comparable medium-sized offices can be credible even when the available data cannot support an equally credible statement about the entire historical stock.

Uncertainty reporting should distinguish sampling error, model uncertainty, and uncertainty in recorded characteristics. Confidence intervals around a fitted cohort contrast address only part of that problem, especially where the sample is policy-defined or inputs are inaccurate. Repeated observations on the same building also require treatment of within-building dependence, while multiple cohort comparisons make isolated significance tests easy to overinterpret. The proposed analysis should therefore report effect magnitudes and uncertainty intervals, explain the dependence structure, and show sensitivity to key definitions. A narrow interval is persuasive only about the statistical quantity actually estimated; it cannot resolve a mismatched boundary, an omitted operational difference, or an unsupported causal interpretation.

Longitudinal records offer a useful way to separate persistent differences from changes in measured operation. They can reveal whether a cohort's observed disadvantage persists across weather years or changes after documented interventions. However, building fixed effects absorb a fixed construction-cohort indicator, and a common linear age trend cannot be separated from calendar time once building-specific constants are included. The protocol must therefore resist claiming that a panel automatically identifies an independent ageing effect. Its strongest immediate contribution is to study within-building changes and the stability of cohort descriptions under consistent measurement. Any further separation of age, period, and cohort requires explicit identifying assumptions or an external source of variation.

Causal evaluation should begin with a specific intervention and an identifiable exposure date. A building-code analysis, for example, would require information about the applicable requirement, permit timing, exemptions, implementation, and a credible comparison group. Completion year alone may misclassify exposure, and a discontinuity around a calendar boundary may coincide with other market changes. Where suitable evidence exists, a design based on a policy change or retrofit timing can be considered, but its assumptions must be examined directly. The synthesis is to reserve causal language for an intervention-specific comparison, while retaining construction-period associations as informative descriptions when those additional conditions cannot be established.

Validation should also reflect the intended application. Holding out records from the same building in different years can overstate predictive performance when the model has already learned persistent characteristics of that property. Conversely, holding out whole buildings or cities asks harder questions about generalization and may reveal that a locally useful cohort relationship does not transfer. A study should therefore align its validation split with the decision it intends to support and distinguish prediction of known properties from transfer to new stocks. Good validation establishes useful predictive scope; it still does not determine whether construction period is a manipulable cause of the predicted energy demand.

Table 2 links the analytical target to an appropriate output. The distinctions are deliberately practical: they indicate what a reader can conclude from each result and what additional evidence is needed to go further. Although a single project may pursue every target, combining them into one ranking would hide their different assumptions. A defensible study should instead allow results to coexist, including cases where the cohort with the highest aggregate demand is not the cohort with the poorest conditional performance. That distinction is central to matching a benchmarking conclusion to an appropriate policy response.

Table 2. Analytical targets and defensible interpretations

Target	Appropriate output	Interpretation boundary
Reported stock burden	Total energy, area-weighted intensity, fuel totals	Describes the covered stock
Typical office	Median and building-weighted distribution	Does not measure total stock burden
Comparable service performance	Conditional contrast with overlap checks	Depends on adjustment choices
Intervention effect	Identified code or retrofit comparison	Requires exposure and causal assumptions

7. Interpretation and implications for office energy policy

Construction period can support initial targeting when detailed technical information is scarce. A cohort associated with particular façade arrangements or servicing systems may warrant a focused audit programme, and a cohort accounting for substantial demand may merit attention even when its typical intensity is unexceptional. However, the same evidence cannot justify assuming that every member requires the same intervention. The practical synthesis is a staged process: use cohort information to organize investigation, verify the relevant physical and operational conditions, and select measures at the building level. This retains the administrative usefulness of construction dates while avoiding expenditure based on an average that may poorly describe the property under consideration.

Operational diagnosis should remain central to that process because a building's annual demand can be strongly affected by when and how equipment runs. Masoso and Grobler (2010) documented substantial consumption outside

working hours in six commercial buildings in Botswana and South Africa. The small, climatically specific sample does not establish a universal avoidable fraction, and some after-hours consumption serves legitimate functions. Its relevance is diagnostic: identical annual totals can conceal very different load schedules and opportunities. A construction-period label cannot reveal whether poor scheduling, continuous equipment loads, or necessary services dominate a particular office, so interval data and an operational review should inform intervention decisions wherever available.

The design-to-operation relationship requires similar care. Menezes et al. (2012) showed, in a central London office case, that more realistic occupancy and equipment information improved predictions of tenant electricity consumption. That result does not quantify a general performance gap across construction cohorts or prove that all design calculations are unreliable. It instead demonstrates why a design expectation and an operational benchmark need aligned boundaries and operating assumptions. For policy evaluation, newer buildings should be assessed against the services and conditions actually delivered, while design compliance remains a distinct question. Confusing those assessments could incorrectly assign an operational problem to the technical standard under which a building was designed.

An illustrative comparison makes the distinction concrete. Consider an older office that has received efficient equipment but remains lightly occupied, alongside a recent office operating longer hours with more intensive tenant equipment. The older property could have lower annual intensity even if the newer property's systems deliver each unit of service more efficiently. Conversely, a favourable service-adjusted result would not erase the newer building's larger actual energy burden. This is a conceptual example, not a finding from a measured pair of buildings. Its implication is that environmental accounting and operational diagnosis should be reported together: one records the demand that must be supplied, while the other helps explain which aspects are amenable to improvement.

For building codes, the comparative evidence supports closer examination of realized outcomes without supporting a blanket verdict based on vintage correlations. Requirements can reduce a component of demand even while other loads or service levels increase, and observed compliance may differ from nominal applicability. At the same time, a policy concerned with absolute energy demand cannot disregard an increase simply because component efficiency improved. Evaluation should therefore distinguish technical effects, implementation, and whole-building outcomes. A credible assessment asks whether a specified requirement changed the outcome it targeted and whether complementary operational or demand measures are needed to achieve the wider objective. That is more informative than treating a cohort slope as a pass-or-fail test of regulation.

Performance normalization also creates a policy choice about which differences should be accepted as legitimate service requirements. Adjusting for operating hours can make comparisons fairer between offices providing different schedules, but it can also remove from the benchmark a dimension of demand that managers can sometimes change. Similar tensions arise for equipment density and conditioned area. The answer is not to abandon adjustment; it is to make its purpose visible. A peer-performance score can account for specified services, while an absolute consumption indicator retains the environmental consequences of providing them. Presenting both avoids making a statistical normalization rule stand in for an unstated judgment about acceptable demand.

Age-based exemptions or requirements would need a stronger justification than a broad cohort average. Older buildings may face particular physical constraints, while recent buildings may offer other opportunities, but the distribution within each group can be wide. Uniform treatment can consequently protect a poorly performing property or burden an already efficient one without improving targeting. A more defensible approach would connect requirements or assistance to verified technical constraints, measured performance, and feasible interventions, using construction period as supporting information. This conclusion follows from the distinction between a useful group descriptor and an individual decision rule; it does not prescribe a

particular legal standard or imply that every jurisdiction has equivalent administrative capacity.

Benchmarking information also cannot, by itself, establish whether replacement is preferable to retaining and upgrading an office. Operational records describe recurring energy within the measured boundary, whereas replacement decisions involve additional consequences, including material production, construction, useful life, and changes in delivered services. An old cohort's high operational intensity may indicate substantial improvement potential, but it is not a complete comparison of alternative life-cycle pathways. Equally, low present energy use does not establish that every retention option is optimal. The defensible conclusion is to use benchmarking to characterize the operational part of the decision and combine it with a separate, explicitly bounded assessment of the alternatives before drawing a replacement recommendation.

Data publication can improve the value of construction-period analysis without requiring a single universal benchmark. Stable property identifiers, clear boundary descriptions, revision histories, and dated records of major interventions would permit researchers to distinguish a physical change from a reporting change. More disclosure is not automatically better, however, when privacy restrictions, reporting burden, or inconsistent field definitions reduce reliability. A useful publication strategy should prioritize variables that resolve specific ambiguities in the energy comparison and provide documentation explaining their meaning. The resulting value comes from reliable connections among records, rather than simply increasing the number of columns available to an analyst.

8. Testable propositions for further research

A first proposition concerns metric-dependent cohort rankings. Where construction periods differ substantially in fuel mix, a ranking based on delivered energy may differ from one based on source energy or historical carbon factors. This is a testable implication of the accounting framework, not a claim that every dataset must exhibit a reversal. A future study should calculate the outcomes from the same records, keep the floor-area boundary constant, and identify which fuel components account for any change in ordering. If rankings remain stable across defensible boundaries, confidence in the descriptive ordering increases. If they change, the study should explain the relevant policy objective rather than select the metric that most strongly supports a preferred historical narrative.

A second proposition concerns service intensity. If part of a raw cohort association reflects operating hours, occupancy, or intensive tenant uses, conditioning on well-measured service characteristics should change the comparison. Yet attenuation alone would not prove that these variables mediate a causal vintage effect, especially if they are measured imperfectly or related to unobserved selection. The appropriate test is therefore a transparent sequence of descriptive and conditional estimates, accompanied by sensitivity to measurement and overlap. The result would establish how much the observed ranking depends on specified service comparisons. A causal explanation would still require stronger assumptions or additional evidence about the relationships among historical design, tenant selection, and operation.

A third proposition concerns the usefulness of renovation history. Adding dated, system-specific interventions may improve prediction and explain some variation within a construction cohort, because original completion year does not describe all currently installed equipment. However, renovated properties may differ systematically from unrenovated properties in ownership, investment capacity, or pre-existing performance. The empirical task is consequently twofold: evaluate whether renovation information improves a clearly defined descriptive or predictive model, and separately investigate intervention effects where the design permits. A positive predictive contribution would justify better data collection even if the records remained insufficient to estimate unbiased retrofit savings. This distinction prevents an operationally useful variable from being assigned more causal meaning than it can support.

A fourth proposition concerns transportability. A construction-period association may weaken when transferred to a city whose building forms, climates, or reporting rules differ from those of the original study. Such weakening would challenge a

universal ranking, although it would not invalidate the local model for its original purpose. Future work should therefore compare locally defined cohorts first, then test a documented harmonization of historical periods and operational boundaries. Transfer should be evaluated on held-out locations, with failures traced to observable differences where possible. The value of this exercise is to establish where a relationship travels and where local evidence remains necessary, rather than to force diverse building histories into a common chronological scale.

A fifth proposition concerns selection into the observed stock. Older buildings that remain offices may differ from those demolished, converted, or no longer reported, while successful redevelopment may change both their use and their metering identity. Surviving old offices could consequently look efficient for reasons that are not shared by the original historical cohort. This is a plausible selection mechanism, not a documented direction of bias in every city. A stronger investigation would link reporting histories with changes in use, demolition, and property boundaries where lawful and feasible. Even incomplete linkage could identify which conclusions concern surviving offices and which claims about historical stock development require a broader population.

9. Limitations of the synthesis

The review provides a structured interpretation of selected evidence rather than an exhaustive map of the global literature. Its principal direct office comparisons concern New York City, England and Wales, and Seoul, and they do not represent all climates, market segments, or construction traditions. This geographical concentration limits the universality of substantive conclusions, although it is sufficient to demonstrate that multiple observed patterns exist. The framework should therefore be judged by whether it clarifies the questions and assumptions of a comparison, not by whether these selected cities establish an international average effect. Additional evidence could refine or challenge particular mechanisms without removing the need to define population and metric consistently.

Access and reporting constraints also limit the depth of cross-study comparison. Some sources provide extensive methodological detail, whereas others make only abstracts or selected sections readily accessible; the article restricts their use accordingly. Even full texts do not necessarily report every distribution, missing-data pattern, or alternative specification needed for direct replication. The synthesis therefore cannot assign a single numerical share of observed cohort differences to morphology, operation, or refurbishment. Its contribution is to identify those unresolved components and prevent unsupported attribution. Resolving them requires more complete reporting or record-level research, rather than additional confidence in a qualitative explanation that the available evidence cannot discriminate.

The publication cutoff creates a deliberately historical evidence base. The article uses material published before 1 May 2023, including dated documentation for survey releases and benchmarking models, so it can assess the knowledge available within that boundary. Later changes in building operation, technology, electricity supply, reporting coverage, or scoring methods could affect a contemporary application. The resulting limitation is temporal applicability, not the validity of accurately reported historical findings. Any future use of the framework should update the underlying records and definitions while preserving the distinction between consumption year, publication year, and model version. Otherwise, a historical comparison could be presented as a statement about conditions it never observed.

The proposed protocol is more demanding than many available datasets can support. Detailed service measures, refurbishment dates, and stable property histories may be incomplete, and a comprehensive design can reduce the number of directly comparable offices. Yet the absence of ideal information does not make benchmarking uninformative. The appropriate response is to narrow the claim to the population, boundary, and target that the records can sustain, while reporting sensitivity where meaningful. A transparent descriptive comparison can support useful investigation even when causal identification is unattainable. The central limitation is therefore not

that every study must answer less, but that each conclusion must stop where its evidential basis stops.

10. Conclusion

Construction period has explanatory value because office buildings retain features of the historical conditions under which they were produced. The reviewed benchmarking evidence nevertheless shows that those historical conditions do not translate into a universal monotonic ranking of operational energy use. Opposing fuel trends, nonmonotonic cohort patterns, and context-dependent associations require a more precise interpretation. Construction period is best understood as a historically structured descriptor whose meaning depends on the measured outcome, the surviving and reporting population, and the services supplied. That interpretation preserves its usefulness while preventing a date from being mistaken for a direct measure of current technical efficiency.

The comparative framework developed here separates stock energy burden, conditional peer performance, and causal intervention effects. These questions can inform one another, but they require different weighting, adjustment, and evidence. Their separation allows a study to report actual environmental demand alongside operational diagnosis without claiming that a correlation identifies what a code change or retrofit would accomplish. The resulting research priority is to connect reliable energy and area records with explicit boundaries, dated interventions, and meaningful service information. Where those connections remain incomplete, construction-period comparisons should remain descriptive or conditional rather than being promoted into unsupported accounts of causation.

For practice, the article supports using construction cohorts to organize audits, investigate historical building characteristics, and identify where better information could improve intervention choices. It does not support uniform performance judgments about individual offices or replacement decisions based only on age. The most useful benchmark is therefore one whose purpose and limits are visible to its users. A construction-period comparison becomes valuable when it explains which demand is being counted, which offices are genuinely comparable, and which next investigation or intervention the evidence can justify. Its contribution is the quality of that decision, rather than the apparent simplicity of an old-versus-new ranking.

References

Ahn, K. U., Shin, H. S., Park, C. S., & Kim, K. W. (2017). Big-data analysis on energy consumption of office buildings in Seoul, Korea. In Proceedings of Building Simulation 2017: 15th Conference of IBPSA (pp. 1540-1547). International Building Performance Simulation Association. https://doi.org/10.26868/25222708.2017.411

Andrews, C. J., & Krogmann, U. (2009). Technology diffusion and energy intensity in US commercial buildings. Energy Policy, 37(2), 541-553. https://doi.org/10.1016/j.enpol.2008.09.085

Armitage, P., Godoy-Shimizu, D., Steemers, K., & Chenvidyakarn, T. (2015). Using Display Energy Certificates to quantify public sector office energy consumption. Building Research & Information, 43(6), 691-709. https://doi.org/10.1080/09613218.2014.975416

Chung, W. (2011). Review of building energy-use performance benchmarking methodologies. Applied Energy, 88(5), 1470-1479. https://doi.org/10.1016/j.apenergy.2010.11.022

Godoy-Shimizu, D., Steadman, P., Hamilton, I., Donn, M., Evans, S., Moreno, G., & Shayesteh, H. (2018). Energy use and height in office buildings. Building Research & Information, 46(8), 845-863. https://doi.org/10.1080/09613218.2018.1479927

Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. CHANCE, 32(1), 42-49. https://doi.org/10.1080/09332480.2019.1579578

Hsu, D. (2014). Improving energy benchmarking with self-reported data. Building Research & Information, 42(5), 641-656. https://doi.org/10.1080/09613218.2014.887612

Hsu, D. (2015). Identifying key variables and interactions in statistical models of building energy consumption using regularization. Energy, 83, 144-155. https://doi.org/10.1016/j.energy.2015.02.008

Kontokosta, C. E. (2015). A market-specific methodology for a commercial building energy performance index. The Journal of Real Estate

Finance and Economics, 51(2), 288-316. https://doi.org/10.1007/s11146-014-9481-0

Kontokosta, C. E., & Tull, C. (2017). A data-driven predictive model of city-scale energy use in buildings. Applied Energy, 197, 303-317. https://doi.org/10.1016/j.apenergy.2017.04.005

Masoso, O. T., & Grobler, L. J. (2010). The dark side of occupants' behaviour on building energy use. Energy and Buildings, 42(2), 173-177. https://doi.org/10.1016/j.enbuild.2009.08.009

Menezes, A. C., Cripps, A., Bouchlaghem, D., & Buswell, R. (2012). Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap. Applied Energy, 97, 355-364. https://doi.org/10.1016/j.apenergy.2011.11.075

Mills, E. (2011). Building commissioning: A golden opportunity for reducing energy costs and greenhouse gas emissions in the United States. Energy Efficiency, 4(2), 145-173. https://doi.org/10.1007/s12053-011-9116-8

Newsham, G. R., Mancini, S., & Birt, B. J. (2009). Do LEED-certified buildings save energy? Yes, but… Energy and Buildings, 41(8), 897-905. https://doi.org/10.1016/j.enbuild.2009.03.014

Scofield, J. H. (2009). Do LEED-certified buildings save energy? Not really… Energy and Buildings, 41(12), 1386-1390. https://doi.org/10.1016/j.enbuild.2009.08.006

Urban Green Council. (2016, August). New York City's energy and water use 2013 report. https://www.nyc.gov/html/gbee/downloads/pdf/nyc_energy_water_use_2013_report_final.pdf

U.S. Energy Information Administration. (2022, December). Table C4. Sum of major fuels consumption and expenditure gross energy intensities, 2018. https://www.eia.gov/consumption/commercial/data/2018/ce/pdf/c4.pdf

U.S. Energy Information Administration. (2023, January). Commercial Buildings Energy Consumption Survey (CBECS): User's guide to the public use microdata file. https://www.eia.gov/consumption/commercial/data/2018/pdf/Users%20Guide%20to%20the%202018%20CBECS%20Public%20Use%20Microdata%20File.pdf

U.S. Environmental Protection Agency. (2019a, July). Analysis and key findings from EPA's review of the ENERGY STAR model for U.S. office properties [Technical reference]. https://www.energystar.gov/sites/default/files/tools/Office-Analysis-KeyFindings_Final508.pdf

U.S. Environmental Protection Agency. (2019b, July). ENERGY STAR score for offices in the United States [Technical reference]. https://www.energystar.gov/sites/default/files/tools/Office_August_2019_508.pdf

U.S. Environmental Protection Agency. (2020, August). Climate and weather [Technical reference]. https://www.energystar.gov/sites/default/files/tools/Climate_and_Weather_2020_508.pdf