Measuring Local Labor Markets with Naturally Occurring Job Ads Data

Meshal Alkhowaiter

London School of Economics and Political Science

Prospectus submitted to the NBER EMRI Conference on Innovations in Measuring Employment

Abstract

This prospectus outlines a methodology for constructing high-frequency, and sub-city labor market indicators by harmonizing microdata from multiple commercial job platforms. I propose a reproducible pipeline that integrates data from platforms such as Glassdoor, Indeed, and regional job boards to produce three classes of indicators currently absent from official statistics: 1) median posted wages by occupation at the sub-city level, 2) job vacancy duration measures using JOLTS-consistent definitions, and 3) geographic job-demand maps identifying emergent employment clusters within metropolitan areas. This prospectus discusses the potential applications to US metropolitan labor markets, and outlines the work planned for presentation at the September 2026 conference.

1. The Measurement gap

The principal US source for job vacancy data, which is the JOLTS program, provides monthly estimates of openings, hires, and separations at the national and regional level, and those statistics are disaggregated by broad industry (Davis, Faberman, and Haltiwanger, 2013). In this paper, I argue that these statistics are indispensable for macroeconomic monitoring, but are also too coarse for local labor market analysis. More specifically, the JOLTS publishes nothing below the census region, offers limited occupational detail, and reports no posted wages. The OEWS program offers occupation-level wage estimates, but those are only available at the metropolitan level on an annual cycle. More broadly, a study by Einav and Levin (2014) shows that large-scale naturally occurring datasets have the potential to transform economic measurement, yet the integration of such data into official statistical programs remains limited, particularly with labor market statistics. However, neither data source captures within-city spatial variation, and whether jobs are clustering in central business districts, dispersing to suburban corridors, or forming new hubs around transport nodes in a given city.

Commercial job platforms have become the dominant channel for posting and finding employment. The Hershbein and Kahn (2018) study shows that job ads data from Burning Glass data could measure employer skill requirements at a large scale, while Cajner et al. (2020) demonstrates that ADP payroll microdata could improve the timeliness and granularity of employment statistics compared to conventional survey-based sources. Similarly, the Marinescu and Wolthoff (2020) study utilizes data from CareerBuilder to analyze how firms determine the wages in job postings, and demonstrates that online vacancy data contains rich information on employer wage-setting behavior that is absent from existing household or establishment surveys. Yet the measurement potential of multi-platform job board data, and particularly for sub-city indicators with no official counterpart, remains largely understudied. A main challenge is the absence of a standardized methodology for integrating data across platforms with different schemas and reporting conventions.

2. Data Sources and harmonization

This paper develops a reproducible pipeline for ingesting raw job-listing data from multiple platforms and outputting a unified analytical dataset. The pipeline addresses four measurement challenges. The **first** challenge revolves around schema unification**,** where each platform structures its data differently. Moreover, the pipeline maps platform-specific fields into a common 15-column schema covering source, city, job title, employer, economic sector, contract type, neighborhood, salary (raw string plus parsed minimum, maximum, and midpoint), years of experience, vacancy duration, listing status, and collection date. **Second**, is economic sector classification**,** where platform-native labels are inconsistent across sources. For instance, Glassdoor uses 17 sector categories plus 33 industry subcategories, while regional platforms may use entirely different taxonomies or no sector field at all. The pipeline implements a two-stage system, where 1) a deterministic lookup table maps known labels to 16 unified sectors, which is

then followed by 2) a keyword-based re-classifier that assigns the remaining listings by matching occupation terms in job titles and employer names. This crosswalk approach draws on methods employed in harmonizing international occupation classifications (ILO, 2012), and achieves economic sectoral assignment for over 90 percent of listings. The third measurement challenge pertains to wage standardization. Specifically, the salary formats vary across platforms, with some sites showing monthly ranges, annual figures, single-point estimates with multiplier notation. The pipeline parses all of these formats into a consistent monthly minimum, maximum, and midpoint, while flagging unrealistic wage values below a jurisdiction-specific minimum wage floor, and preserves the raw string for audit, which parallels the methods implemented by Autor, Katz, and Kearney (2008) when harmonizing wage data across CPS redesigns. The fourth key challenge relates to deduplication**.** This occurs because a single vacancy may generate multiple records in the merged dataset through two channels: 1) the same employer posting on different platforms, and 2) repeated captures of the same listing across successive scraping sessions. The pipeline deduplicates within each source on employer-title-location tuples, retaining the most recently posted instance to preserve the freshest salary and duration information. After merging across sources, a second deduplication pass on the same tuple structure removes any cross-platform overlaps. This matters for measurement because unresolved duplicates would inflate vacancy counts and bias the demand concentration estimates the pipeline is designed to produce.

3. Proposed Indicators for US Metropolitan Labor Markets

The methodology outlined above can be employed to improve labor market data measurement and collection, where websites such as Indeed, Glassdoor, LinkedIn, and ZipRecruiter collectively capture a substantial and growing share of job ads, and especially in urban and suburban areas. More importantly, these platforms contain structured fields for occupation, geographic location often at the ZIP code level, posted salary ranges, and employer identifiers, which I believe are precisely the variables needed to construct labor market indicators at a granular level that existing labor force surveys may not capture.

The first proposed indicator is a monthly index of median posted wages by detailed occupation and ZIP code. The Occupational Employment and Wage Statistics (OEWS) program currently produces the closest official equivalent, but those estimates are published annually and only at the metropolitan area level. However, a job-board-derived wage index would offer three measurement advantages. **First**, higher temporal frequency, which enables researchers and policymakers to observe wage adjustments in near-real time. **Second**, it offers a finer geographic resolution that allows for comparisons across neighborhoods and commuting zones within the same metro area. This is particularly valuable for understanding spatial wage gradients that may reflect differences in local cost of living, employer concentration, as well as proximity to employment clusters. This spatial dimension is particularly useful given the findings of Manning and Petrongolo (2017), where the authors show that job seekers search locally and that labor

markets are considerably more geographically segmented than aggregate statistics may otherwise suggest. **Third**, because job-board wages tell us about what companies are offering in the private sector, rather than what workers are currently earning, comparing the two series opens a measurement dimension that neither source can provide by itself. For example, a persistent gap between 1) posted wages and 2) OEWS-reported actual wages in a given occupation or area may signal that employers are bidding up offers to attract scarce workers, or conversely, if the advertised salaries systematically overstate actual salary levels. Specifically, tracking this offered-versus-paid wage differential across time would provide a leading indicator of wage pressure that could complement the lagging measures available from employer surveys. This distinction between posted and realized wages also connects to the wage-posting literature (e.g. see Hall and Krueger, 2012), which documents how a substantial share of US employers post non-negotiable wage offers, and thus providing evidence that the advertised figure is a meaningful economic variable rather than merely an opening bid.

The second indicator is a vacancy duration distribution by occupation and metropolitan area. Although the JOLTS program reports monthly counts of job openings, it does not track how long individual positions remain posted. Therefore, by calculating the elapsed time between a job listing's first appearance and its removal or expiration, the pipeline offers policymakers a continuous duration measure that can be aggregated into distributional statistics. Those may include 1) median days to fill a job in a given sector or occupation, and 2) the share of listings exceeding 30 or 60 days. These duration descriptive indicators would allow researchers to identify hard-to-fill occupations before they appear in the aggregate labor shortage indicators, which offers an early signal of emerging labor market tightness in specific geographic areas or economic sectors.

The third indicator is a sub-metropolitan demand map that tracks the geographic distribution of job postings, which essentially captures labor demand within cities throughout time. More specifically, by geocoding job advertisements on sites such as Indeed to ZIP codes or neighborhoods, the pipeline can identify emergent employment hubs, measure the degree of spatial concentration in particular economic sectors, and monitor whether job demand is shifting toward or away from central business districts. I believe that this type of employment data addresses a longstanding blind spot in official employment data, where the Bureau of Labor Statistics produces metro-level estimates, no existing federal program measures where within a metropolitan area jobs are being created on a recurring basis. For urban labor economists, this indicator is valuable because detailed information on spatial distribution of vacancies has direct implications for commuting costs, local unemployment disparities, and the effectiveness of place-based employment policies. Marinescu and Rathelot (2018) offer empirical evidence demonstrating that the geographic distance is a major determinant of job application behavior amongst unemployed individuals, which further implies that the spatial distribution of vacancies within a city has direct consequences for which workers can access which jobs, and further reinforces the case for sub-metropolitan labor market measurements. Similarly for metropolitan

planning agencies, it offers an empirical basis for identifying which neighborhoods are emerging as employment clusters and which are losing private firm activity, which is information that currently relies on infrequent census surveys or ad hoc local surveys, rather than systematic and high-frequency measurement.

An important consideration in constructing these indicators is the data representativeness, which should be acknowledged clearly. Not all employers or occupations post on commercial job boards at equal rates, and the resulting statistics must be interpreted with these coverage patterns in mind. For example, certain economic sectors such as technology and professional services, are likely to be overrepresented in online job ads compared to industries where hiring occurs more through informal channels or internal referrals such as restaurants. The planned validation exercise described below is designed in part to quantify these coverage gaps by comparing the employment data derived from online platforms against their closest official counterparts.

4. Current status and future steps

As of April 2026, the core pipeline architecture is fully developed and operational. The schema unification, two-stage sector classification, salary parsing, and deduplication modules have been built and tested across three structurally distinct platforms, including sources with different languages, taxonomies, and salary reporting formats. The analytical framework for producing all three indicator classes, which are 1) wage distributions, 2) vacancy duration measures, and 3) geographic demand maps is complete and documented with reproducible code on R studio.

Between now and the September conference, I intend to extend this work in two directions. First, I will adapt the pipeline to US data by acquiring licensed bulk access to job-listing microdata from one or two major platforms for two to three metropolitan areas. I believe the harmonization architecture is well suited for this, as it is platform-agnostic by design, and adapting it to a new data source requires only the addition of source-specific lookup tables for field mapping and economic sector crosswalks, but the core logic for salary parsing and deduplication remains unchanged. Second, I intend to benchmark the resulting indicators against their closest published BLS and other surveys. More specifically, the advertised wage distributions will be compared to OEWS metropolitan-level estimates for overlapping occupations, and the vacancy duration shares will be compared to those in the JOLTS regional aggregates. I think that this validation exercise serves a dual purpose where 1) it establishes where the naturally occurring data complement official statistics, and 2) it also identifies the conditions under which the data coverage gaps may produce systematic divergence from published official statistics.

In conclusion, the conference presentation would cover four main sections:

1. The measurement framework and pipeline architecture,

2. The indicator construction methodology for wage, job vacancy duration, and spatial demand measures,
3. initial US results for the selected metropolitan areas together with the validation benchmarks against OEWS and JOLTS publications, and
4. A discussion of the data representativeness, coverage limitations, and the practical conditions under which online job ads can reliably supplement official employment statistics.

Conflict of Interest Statement. The author has no financial or other interests related to the commercial job platforms discussed in this prospectus. This research has not received funding from any job platform company or data vendor.

References

Abraham, K. G., Haltiwanger, J. C., Sandusky, K., and Spletzer, J. R. (2021). "Measuring the Gig Economy: Current Knowledge and Open Issues." In C. Corrado, J. Haskel, J. Miranda, and D. Sichel (Eds.), Measuring and Accounting for Innovation in the Twenty-First Century. University of Chicago Press.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008). "Trends in U.S. Wage Inequality: Revising the Revisionists." Review of Economics and Statistics, 90(2), 300–323.

Cajner, T., Crane, L. D., Decker, R. A., Grigsby, J., Hamins-Puertolas, A., Hurst, E., Kurz, C., and Yildirmaz, A. (2020). "The U.S. Labor Market During the Beginning of the Pandemic Recession." Brookings Papers on Economic Activity, 2020(2), 3–34.

Davis, S. J., Faberman, R. J., and Haltiwanger, J. C. (2013). "The Establishment-Level Behavior of Vacancies and Hiring." Quarterly Journal of Economics, 128(2), 581–622.

Einav, L. and Levin, J. (2014). "Economics in the Age of Big Data." Science, 346(6210), 1243089.

Hall, R. E. and Krueger, A. B. (2012). "Evidence on the Incidence of Wage Posting, Wage Bargaining, and On-the-Job Search." American Economic Journal: Macroeconomics, 4(4), 56–67.

Hershbein, B. and Kahn, L. B. (2018). "Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings." American Economic Review, 108(7), 1737–1772.

International Labour Organization (2012). International Standard Classification of Occupations: ISCO-08. Geneva: ILO.

Manning, A. and Petrongolo, B. (2017). "How Local Are Labor Markets? Evidence from a Spatial Job Search Model." American Economic Review, 107(10), 2877–2907.

Marinescu, I. and Rathelot, R. (2018). "Mismatch Unemployment and the Geography of Job Search." American Economic Journal: Macroeconomics, 10(3), 42–70.

Marinescu, I. and Wolthoff, R. (2020). "Opening the Black Box of the Matching Function: The Power of Words." Journal of Labor Economics, 38(2), 535–568.