第 11 回 ビッグデータ等の利活用推進に関する産官学協議のための連携会議
議事概要

１　日　時　　令和 2 年 12 月 23（水）15:00～17:00

２　場　所　　総務省第二庁舎　６階特別会議室（Web 会議）

３　出席者

・ 構成員　　高橋座長、庄司構成員、田原構成員、水野構成員
　　清家構成員（オブザーバー）
・ 審議協力者　横浜市立大学　佐藤教授
　　東京大学　竹内教授
　　ＪＡＸＡ　石田特任担当役
　　株式会社リクルートキャリア　経営統括室　高田悠矢氏
・ 事務局　　総務省政策統括官（統計基準担当）付　統計委員会担当室

４　議　題

（１）観測データ利活用検証 WG の中間報告
（２）労働ビッグデータの統計的利活用について
（３）ビッグデータ連携会議におけるこれまでの事例整理について

５　配付資料

資料1 SDG15.4.2（山地緑地被覆指数：MGCI）検証作業報告
資料2 地球観測データを用いた SDG15.4.2（山地植生被覆指数）の試算について（中間報告）
資料3 労働市場のビッグデータ：経済統計としての活用の可能性③
資料4 公的統計へのビッグデータの更なる活用に向けて
資料5 公的統計へのビッグデータの更なる活用に向けて（概要版）
参考資料　第 10 回ビッグデータ等の利活用推進に関する産官学協議のための連携会議議事概要

６　議事概要

（１）観測データ利活用検証 WG の中間報告

概要は以下の通り。

○　横浜市立大学・佐藤教授より資料１、JAXA 石田特任担当役より資料 2 の説明が行われた。

主な質問・意見は次のとおり。

○　FAO の MGCI 試算値と JAXA の MGCI 試算値で乖離がある要因は、FAO の値は ESA の被覆データを基にして計算され、その被覆の分類がそもそも実際と異なるという理解でよろしいのか。

- ヨーロッパの ESA の 300 メートルの土地被覆データと、ＪＡＸＡの高解像度の土地被覆データの違いによる乖離です。

○　佐藤先生から KAPOS2、3 について説明がありましたが、スライドを拝見すると、KAPOS4、5、6 についてもやはり FAO と乖離しているように見えますが、全体的に過大に推計されているということか。

- KAPOS4、5、6 については、FAO、KAPOS4 では 99.68 というのを出していますが、JAXA では 96.4 となっていますが、まだ未検証であり、今後取り組ませていただきたいと思っております。

（２）労働ビッグデータの統計的利活用について

概要は以下の通り。

○　株式会社リクルートキャリア　経営統括室　高田氏より資料3の説明が行われた。

主な質問・意見は次のとおり。

○　このデータ自体貴重なデータと感じ、政府統計を無理に当てようとしなくても、そのままこのデータを指標化して、雇用動向調査とは別の統計として公表するだけでも価値はあるし、利用しやすいと思う。その参考系列として、割合単純な膨らませ方で雇用動向調査の母集団に合わせて膨らませたらこのようになるというのを出すだけでも、十分に価値のある情報だと思った。

○　３つ推計手法を試されていて、手法すべてに対し属性がたくさんあるようだが、属性ごとに、リクルートキャリアのデータと雇用動向調査で賃金上昇の割合がどれぐらいの差があるのかは比較されていると思うが、属性によって賃金上昇割合は連動しているものなのか？また、属性は全部使われているのか、どの属性が重要か、属

性の取捨選択をしているのかを伺いたい。あまり属性が多過ぎると、オーバーフィットが起きたりしないか？

- 属性で捉えられない部分もある程度存在する。結果として連動しない部分は過去データに基づき、定数を乗じることで補正する事で対処をしている。
- リクルートキャリア側のデータはかなり多くの属性情報が存在するが、これらの推計では雇用動向調査にも存在する属性情報以外は利用できない。雇用動向調査は企業へのアンケートをもとにしているため、設問数、即ち取得できる属性情報はそれほど多くない。故に、ここでは物凄く多くの属性情報を用いているというような状況ではそもそもないが、変数選択を行っているのかという意味で質問を捉えると、今回は Stepwise のような探索は行わず、Elastic net による正則化法を用いているという回答になる。

○ 推計手法の２つ目と３つ目で、２つ目は分類で３つ目は回帰の問題として解いたと説明があったが、３つ目の回帰は何割上がったかという上昇率を左辺に持ってきたということなのか、その場合、どのように説明資料のＹに膨らましたか教えていただきたい。

- 分類問題のケースではシンプルにＹとして、賃金が 1 割上がったか否か、1 か 0 の 2 値変数を用いて学習している。結果として算出されるスコアは 0 から 1 の連続値となり、これを「転職時に賃金が 1 割以上増加する確率」と見做して利用している。回帰問題のケースでは、代わりにＹとして「前職と次の仕事の賃金の比」を用いており、これらは 1.10 や 1.12 といった連続値である。こちらは結果として算出されるスコアも同様のレンジの連続値となるが、それらの値を 0 から 1 の連続値へ変換し、分類モデルの時と同様に「転職時に賃金が 1 割以上増加する確率」と見做して集計を行っている。

○ 推定したもとの雇用動向調査は調査対象期間が 1 年でデータ数が少ないので、比較することは難しいですが、資料によると、2011 年の上がり幅が少し大きいように、このデータの場合でも傾向が出てきている。これは統計的に発生する揺らぎなのか。それとも、2011 年に関して誤差が生じやすいような属性なり何かの要因というのがあったのか。

- 「雇用動向調査とリクルートキャリアのデータで、時系列の振る舞いが大きく異なる年があるのは何故か」という質問であると解釈すると、現段階においては厳密には分からないというのが正直なところ。ご指摘の 2011 年だけでなく、雇用動向調査では 2016、17 年に 1 回大きな起伏があるが、ここは弊社のデータでは滑らかである。これらの原因を特定するための分析は現段階では行っておらず、一旦は測定誤差のようなものと見做して飲み込んでいる。仮に測定時の誤差でないとすると、特定の属性だけに非常に大きな変動が起こるようなケースか、属性毎の構成比に大きな変化があるケースにおいて、このような事象が起こると想定されるが、いずれにしろ現段階では単なる想像の域

を出ない。この辺りは時間が許せば追加的な検証をしていきたい。

（３）ビッグデータ連携会議におけるこれまでの事例整理について

概要は以下の通り。

○ 総務省統計改革実行推進室谷道企画官より資料４の説明が行われ、ビッグデータ連携会議におけるこれまでの事例整理及びその概要について公表することとされた。

以上