LA EXPLICACIÓN CAUSAL EN ECONOMÍA[*]

Luis Lorente S.-B.[a]

[*] DOI: https://doi.org/10.18601/01245996.v20n39.02. Recepción: 23-01-2018, modificación final: 10-04-2018, aceptación: 11-05-2018. Sugerencia de citación: Lorente S.-B., L. (2018). La explicación causal en economía. *Revista de Economía Institucional*, *20*(39), 9-51.

[a] Profesor Emérito, Facultad de Ciencias Económicas, Universidad Nacional de Colombia.

La explicación causal en economía

Resumen. La teoría económica rehúye las explicaciones de tipo causal y adopta relaciones funcionales y soluciones de equilibrio que suponen alguna forma de simetría entre las variables involucradas. Sin embargo, la causalidad es un supuesto inherente a toda política o decisión económica, que se defiende y adopta solo porque se espera de ella un determinado efecto. La desconfianza proviene de que, en ausencia de un método experimental que aísle unos fenómenos de su entorno y garantice su repetibilidad, es posible imaginar numerosas causas, incluso teleológicas, que interfieren mutuamente y ofrecen un exceso de explicaciones alternativas, sin un criterio claro para escoger entre ellas. Pero existen circunstancias que permiten deducir regularidades de tipo causal en un contexto de cambio dinámico, innovación y desequilibrio permanente. Este artículo desarrolla varios ejemplos sencillos que tienen consecuencias importantes para la teoría económica y sugieren ajustes de fondo en algunas políticas.

Palabras clave: identidad contable, correlación estadística, explicación causal, teoría explicativa, análisis macroeconómico, política económica; JEL: B41, E60, G23, O31.

The causal explanation in economics

Abstract. Economic theory avoids explanations of causal type and adopts functional relationships and equilibrium solutions that suppose some form of symmetry between the variables involved. However, causality is an inherent assumption of any economic policy or decision, which is defended and adopted only because a certain effect is expected from it. The distrust stems from the fact that, in the absence of an experimental method that isolates phenomena from its environment and guarantees its repeatability, it is possible to imagine numerous causes, even teleological, that interfere with each other and offer an excess of alternative explanations, without a clear criterion to choose among them. However, there are circumstances that allow the deduction of regularities of causal type in a context of dynamic change, innovation and permanent imbalance. This article develops several simple examples that have important consequences for economic theory and suggest fundamental adjustments in some policies.

Keywords: Accounting identity, statistical correlation, causal explanation, explanatory theory, macroeconomic analysis, economic policy; JEL: B41, E60, G23, O31.

A explicação causal em economia

Resumo. A teoria econômica evita as explicações de tipo causal e adota relações funcionais e soluções de equilíbrio que supõem alguma forma de simetria entre as variáveis envolvidas. Contudo, a causalidade é um suposto inerente a toda política ou decisão econômica, que é defendida e adotada somente porque se espera dela um determinado efeito. A desconfiança provém de que, na ausência de um método experimental que isole uns fenômenos de seu contexto e garanta sua repetibilidade, é possível imaginar numerosas causas, inclusive teleológicas, que interferem mutuamente e oferecem um excesso de explicações alternativas, sem um critério claro para escolher entre elas. Porém, existem circunstâncias que permitem deduzir regularidades de tipo causal num contexto de mudança dinâmico, inovação e desequilíbrio permanente. Este artigo desenvolve vários exemplos simples que têm consequências importantes para a teoria econômica e sugerem ajustes de fundo em algumas políticas.

Palavras-chaves: identidade contábil, correlação estatística, explicação causal, teoria explicativa, análise macroeconômica, política econômica; JEL: B41, E60, G23, O31.

La causalidad parece ser una palabra proscrita en el lenguaje de la teoría económica actual. Su desprestigio se debe en parte a las críticas de filósofos empiristas iniciadas por Hume y, en especial, de la escuela neopositivista a comienzos del siglo XX. Además, es una consecuencia de la simetría de fuerzas implícita en los análisis usuales de equilibrio económico, cuyo resultado final es una simple relación funcional entre variables, más o menos bien definidas, que juegan papeles de igual magnitud y sentido opuesto.

Esto último parece seguir la tendencia observable en física, donde se prefiere hablar de legalidad –otra forma de describir una relación funcional–, en vez de hablar de causas propiamente dichas. Sin embargo, las dudas de los físicos acerca de la causalidad desaparecen cuando se discute la validez del método experimental, pues esta forma de investigar y validar no tendría justificación alguna salvo que se crea firmemente en que los resultados son consecuencia del experimento y, por ende, causados por una decisión del experimentador.

En general, las ecuaciones de la física relacionan estados, es decir, valores específicos de ciertas variables que describen adecuadamente una situación o un fenómeno concreto, y no podemos decir que un estado sea causa de otro, así como sería erróneo atribuir una propiedad causal a los valores iniciales o a cualquier otro parámetro de una ecuación diferencial.

La causalidad se da entre eventos, es decir, entre cambios de estado: la observación de lo que sucede cuando el experimento altera una variable es el origen de la relación funcional entre ella y otra variable que resume la consecuencia del experimento en cuestión: el método experimental explora el efecto de cambios deliberados.

Después, cuando ya se tienen varias relaciones de este tipo entre variables afines, es posible construir un esquema deductivo que las relaciona entre sí. O bien, cuando el objeto de estudio es un sistema compuesto por partes que interactúan, es posible explicar los cambios del agregado como resultado de cambios sucedidos en el nivel desagregado, y construir entonces una explicación causal que relaciona un nivel con otro, pero siempre referida a cambios de estado.

En este proceso, la construcción de teorías relaciona propiedades y ecuaciones aplicando la lógica deductiva, arriesgando a veces hipótesis que no tienen un origen empírico, o empleando variables que no tienen una contraparte observacional, pero cuyas consecuencias admiten alguna forma de validación mediante la observación o el experimento.

Sin embargo, es vital la diferencia entre (a) la relación funcional y el esquema lógico-deductivo, que son ambos formalismos con fre-

cuencia aplicables a varios fenómenos muy distintos entre sí, y (b) la teoría propiamente dicha, que surge de las interpretaciones atribuidas a esos símbolos formales, se expresa en hipótesis con contenido fáctico y, por consiguiente, atañe a un fenómeno específico. Por ejemplo, las ecuaciones y fórmulas de crecimiento exponencial aparecen en demografía, economía, biología, química y en una infinidad de aplicaciones de la física y la ingeniería, pero, en cada caso, se convierten en una teoría diferente cuando les añadimos una interpretación específica, o unos supuestos causales, u otra forma de explicación de los hechos representados.

A medida que avanza y se enriquece, cada teoría va tomando la forma de un esquema deductivo, pero construido a partir de hipótesis relativamente simples que reflejan un resultado empírico, de observaciones o experimentos que relacionan cambios de ciertas variables o estados con cambios en otras variables o estados. El propósito de esta construcción es explicar unos hechos refiriéndolos a hechos ya conocidos, aunque no se excluye que las adiciones descalifiquen la validez de la construcción teórica anterior y lleven a sustituirla por otras explicaciones más completas o incluyentes.

La construcción de ese andamiaje deductivo incluye relaciones funcionales, muchas veces deducidas con ayuda de razonamientos matemáticos que actúan a manera de argumento lógico prefabricado, y va tomando la apariencia de un sistema descrito por un conjunto de estados y ecuaciones que parece ajeno a todo supuesto causal.

Este es el método que caracteriza el avance tecnológico y científico desde el siglo XVIII, aunque sus orígenes se remontan a Galileo y Newton, y difiere del método aristotélico que predominó en la antigüedad hasta su máximo desarrollo en los escolásticos: un método deductivo que partía de ciertas premisas razonables que, según se creía, expresaban la naturaleza de las cosas y de sus formas de cambio.

No es cierto que los aristotélicos despreciaran la observación empírica y el experimento: la diferencia esencial está en la interpretación de los resultados que, en su caso, solo admitía lo que fuera compatible con sus premisas generales, que no ponían en discusión. En la práctica, ni siquiera podían someter esas premisas a prueba empírica porque solo podían examinar lo que esperaban ver de acuerdo con sus ideas preconcebidas, y porque solo consideraban lícitas las pruebas compatibles con aquellas premisas que habían adoptado *a priori*, por simple *razón pura*.

Por eso mismo, toleraban construcciones eminentemente empíricas, como los epiciclos de la astronomía tolemaica, y admitían que

fueran sustituidas por estructuras muy diferentes, como el sistema copernicano, siempre y cuando se interpretaran como simples construcciones fenomenológicas, como *cajas negras* que describían las apariencias, pero no la esencia, naturaleza o razón de las cosas. Podemos decir que el conflicto cultural no comenzó con Copérnico, que aceptaba el movimiento circular como el único perfecto, sino con las elipses de Kepler, aunque todavía era posible verlas como una simple descripción fenomenológica. La crisis apareció con la interpretación de Galileo, y alcanzó el carácter de revolución mental con Newton.

De ahí en adelante, el método deductivo fue desapareciendo de otras disciplinas, sustituido por lo que se dio en llamar el método científico, que podríamos describir como un camino de ida y vuelta de la observación a la hipótesis; luego, de esta última a la relación con otras observaciones e hipótesis para integrarlas en teorías que permiten deducir nuevas hipótesis, regresando finalmente a la observación y construcción de nuevas y mejores explicaciones.

A lo largo del camino, han sido frecuentes los argumentos del tipo *navaja de Occam* para criticar las nuevas teorías, afirmando que no se deben complicar los modelos ni las teorías ya establecidas para explicar de otra manera los mismos fenómenos. Pero este es un uso inapropiado de Occam: primero, porque toda teoría nueva modifica la visión del problema y puede sugerir observaciones y resultados diferentes, que la visión anterior excluía porque eran contrarios a sus premisas; y, segundo, porque congela la teoría vigente, así como la posición de quienes se oponían a la relatividad porque era más complicada que la mecánica de Newton y coincidía con ella en todos los usos prácticos imaginados hasta ese momento. Si esta postura hubiese prevalecido, jamás habrían aparecido aplicaciones y servicios que hoy son tan comunes como el sistema satelital de posicionamiento global, GPS.

Ciertamente, el argumento de la simplicidad fue uno de los temas favoritos de Einstein cuando intentaba explicar sus criterios de construcción teórica, pero se debe entender como un principio limitado por el tipo de fenómeno y por la clase de explicación que se intenta conseguir: no tendría sentido rechazar el espacio curvo de Riemann solo porque el espacio euclídeo es más simple, ya que en este último no es posible expresar las mismas relaciones que admite aquel.

La búsqueda de explicaciones simples suele desembocar en supuestos de linealidad y no hay duda de que el rápido desarrollo de la física –y la aplicación del método experimental en ella y en otras ciencias– obedeció a que los problemas estudiados eran, casi todos, de carácter lineal, es decir, que admitían la separación en aspectos

relativamente independientes entre sí (fase de análisis), cuyas consecuencias luego podían sumarse para hallar un resultado conjunto (fase de composición o superposición). De ahí que muchos teóricos consideraran la linealidad como una característica esencial e inherente a la causalidad.

Entre ingenieros, médicos y, en general, en las áreas de la tecnología, la práctica cotidiana habla de causas y efectos, aunque no con la sencillez y generalidad que se atribuía a la explicación causal en el siglo XIX, con su proyección simplista de la idea sensorio-motriz abstraída de la experiencia diaria de cada ser humano, porque en la segunda mitad del siglo pasado surgieron novedades que afectan la interpretación y la explicación de los eventos que nos rodean, pero no son compatibles con la noción primitiva de causa.

La primera fue la cibernética, donde las cadenas causales que se cierran sobre sí mismas proporcionan controles automáticos y explican formas de estabilidad dinámica que antes solo cabía atribuir a un comportamiento teleológico.

Al mismo tiempo, con el desarrollo de teorías generales de sistemas, adquirió pleno sentido el emergentismo: la idea de que la organización interna, determinada por las relaciones mutuas entre los componentes del sistema, puede gestar propiedades nuevas en ese agregado, que no están presentes en ninguna de sus partes por separado. La teoría de la complejidad muestra cómo la interacción de componentes que, cada uno por separado, tienen actuaciones muy simples, puede explicar comportamientos extraordinariamente complejos del conjunto, incluso cuando dichas relaciones mutuas son débiles u ocasionales. Esto abre el camino para hablar de un determinismo estadístico, asimilable al resultado de un proceso estocástico.

Algo semejante se percibió en la física con la mecánica estadística, que explica las leyes de la termodinámica clásica como resultado de la interacción desordenada de un inmenso número de moléculas y es un buen ejemplo de explicación de un nivel *macro* recurriendo a elementos de un nivel *micro* inmediato y más simple. Esas leyes de la termodinámica clásica luego aparecen como simples relaciones funcionales entre unas variables macro, aparentemente libres de toda connotación causal. Y, aunque por razones muy diferentes, esa misma apariencia de relación funcional ajena a la causalidad va apareciendo, casi en forma simultánea, en otras ramas de la física, con el desarrollo de las teorías del campo electromagnético y, más tarde, con las del campo gravitatorio. Todas estas novedades de finales del siglo XIX y comienzos del XX ayudan a entender la postura neopositivista que,

por esa época, generaliza las críticas de Hume a la causalidad e intenta sustituirla por la idea de ley natural.

Con el análisis cibernético, la explicación teleológica o finalista parece reducirse al concepto normal de causalidad eficiente, porque si hay un camino causal que conduce al resultado deseado y garantiza su estabilidad, y si ese fin imaginado es una consecuencia de unos motivos que también tienen explicación en la historia pasada de quien decide, no se necesita el futuro para explicar el presente ni el pasado (Nagel, 1961). Pero esta reducción no es total: la unicidad del fenómeno histórico, que no se pliega a leyes universales, depende de una confluencia de condiciones que hacen posible su ocurrencia y que no podemos proyectar como haríamos con una relación funcional: hay un margen de casualidad, de unicidad irrevocable, que se debe a la diferencia esencial entre simple condición contingente y relación legal necesaria o funcional.

Un detalle importante de la explicación sistémica es que recurre a los componentes del nivel inmediato, sin dar saltos de los agregados globales a las unidades más elementales. Por ejemplo, la teoría cinética de gases recurre al movimiento de partículas para explicar la presión, pero se detiene en el nivel molecular sin recurrir al átomo ni a los componentes subatómicos, así como utiliza la mecánica newtoniana de choques elásticos sin recurrir a explicaciones propias del nivel cuántico.

El salto de niveles puede dejar a un lado los rasgos generalizables del fenómeno y trae consigo el riesgo de caer en lo circunstancial y anecdótico. Así, en historia, igual que en economía, debemos reconocer que el nivel inmediato es el grupo social y no el individuo. Por ejemplo, no es posible explicar la influencia del líder sin tomar en cuenta el entorno social que permitió su ascenso, ni la presión de sus seguidores, que a veces lo obligan a ir más allá de sus intenciones originales. De igual manera, tampoco podemos explicar el comportamiento económico del individuo sin tomar en cuenta el del grupo a que pertenece o al que desea adscribirse, porque saltarse ese nivel puede llevarnos a revivir las mónadas de Leibniz o a aceptar su equivalente moderno: el hombre racional de la teoría neoclásica que tiene preferencias prefijadas y actúa con independencia de todos los demás individuos.

En contraste con esta última teoría, a fines del siglo XIX apareció una explicación del consumo que podríamos llamar sistémica, porque atribuye el comportamiento individual a la imitación dentro de un grupo o estrato, combinada con la emulación del estrato superior (Veblen, 1899). Cuatro decenios después se adaptarán estas explicacio-

nes al consumo agregado (Duesenberry, 1949) y, más recientemente, reaparecen con el nombre de "cascadas de consumo" para explicar el comportamiento de los ahorros cuando hay concentración del ingreso (Frank y Levine, 2007).

Después de las teorías de sistemas, y con pocos años de diferencia, surgieron la geometría fractal y la dinámica caótica de sistemas no lineales, las cuales demuestran que sistemas sencillos y perfectamente deterministas pueden llevar a resultados imposibles de prever, incluso indistinguibles de procesos estrictamente estocásticos. Algo tan simple como la multiplicación de números enteros, seguida del truncamiento de las primeras cifras obtenidas, genera números aleatorios con distribución uniforme: basta escoger dos números primos suficientemente grandes, uno como multiplicador y el otro como semilla, para obtener una serie de números que ninguna prueba estadística consigue distinguir de otra serie estrictamente aleatoria, pero que se puede generar cuantas veces se desee sin cambio alguno (Knuth, 1969).

Otro hallazgo de enorme importancia es que muchos sistemas no lineales admiten *atractores extraños*, llamados así porque combinan dos propiedades que, hasta ese momento, se creían antagónicas. Aparecen en sistemas continuos que tienen al menos tres variables de estado, es decir, que son tridimensionales, y son subconjuntos acotados del espacio de estados donde se mueve el sistema. Están contenidos dentro de un volumen finito de posibles puntos de dicho espacio de estados, pero tienen dimensión fractal, así que no llenan el volumen que les sirve como recipiente. Las trayectorias del sistema que llegan a la zona de atracción permanecen luego en dicho atractor, en un movimiento que es recurrente, porque la trayectoria que pasa por un punto vuelve a pasar infinitas veces por otros puntos cercanos, pero que no es periódica porque no se repite jamás: cuando se dibuja la gráfica de una de esas variables de estado contra el tiempo, se ve una serie de segmentos que parecen seguir cierto orden, a veces casi periódicos, que, de pronto, sufren un cambio drástico y comienzan a fluctuar de manera muy distinta. Las ecuaciones del sistema pueden ser estrictamente deterministas, pero la trayectoria no es predecible, salvo en lapsos cortos, aunque tampoco sea posible asegurar cuándo terminará esa relativa predecibilidad y aparecerá otro cambio de marcha drástico. Si comparamos dos trayectorias que comienzan en dos puntos del atractor ligeramente separados, la distancia entre esos puntos, medida a lo largo de la trayectoria que siguen, aumenta exponencialmente hasta que supera el diámetro del volumen que actúa como recipiente del atractor: a partir de ese instante, solo es posible

decir que ambas trayectorias permanecen sobre el mismo atractor, pero saber por dónde va una no puede decirnos nada acerca del lugar por dónde va la otra. Es una situación similar a la de un jugador de billar que lanza con fuerza una bola: si es hábil, puede prever con exactitud la trayectoria después de un par de rebotes, pero después de cuatro o cinco solo podrá decir que la bola debe estar en algún lugar de la mesa, ya que los bordes de ésta no son espejos perfectos y los choques sucesivos amplifican el efecto de cada minúscula diferencia.

Este comportamiento recuerda las características de las fluctuaciones macroeconómicas, que nunca se repiten en forma idéntica, aunque sean semejantes en varios aspectos, porque difieren en otros comportamientos que, además, no son siempre los mismos. Los modelos económicos no lineales pueden presentar esta clase de atractores extraños, aunque tengan pocas variables de estado y parezcan relativamente sencillos, como los modelos Keynes-Metzler-Goodwin (Chiarella et al., 2005). Algunos de ellos son el perfecto ejemplo de un sistema plenamente determinista que admite previsiones de corto plazo pero es impredecible a más largo plazo.

El recuento anterior es importante porque muestra cómo ha cambiado la naturaleza de las explicaciones en la física y, en general, en las áreas técnicas: en vez de un desarrollo lineal y acumulativo, sobrevienen sacudimientos ocasionales que cambian las premisas más fundamentales: las que establecen cuáles son los problemas pertinentes y cuáles respuestas son aceptables. No es un progreso paulatino, basado en la aplicación paciente de un método específico, sino una sucesión de cambios, más o menos drásticos, que modifican las percepciones más fundamentales acerca del objeto de estudio. Cuando esto sucede, cambia también el tipo de teorías explicativas y las nuevas son, muchas veces, más complejas que las anteriores.

LA ECONOMÍA CONVENCIONAL

La economía neoclásica conserva un aire de escolasticismo porque parte de premisas indiscutidas, como cierta versión de la racionalidad humana que traduce en un supuesto universal de optimización, o la idea de que los agregados son una simple suma de comportamientos individuales, cada uno regulado por preferencias fijas e independientes entre sí. Sobre esta base construye un armazón lógico-deductivo, apoyado en modelos matemáticos que acuden, en última instancia, a un supuesto de influencias exógenas para explicar el cambio técnico o las fluctuaciones recurrentes.

Esta clase de teorías parece irrebatible porque sus defensores solo admiten pruebas compatibles con sus premisas: pueden tolerar inconsistencias protuberantes con la esperanza de descubrir algún día argumentos *ad hoc* que resuelvan la contradicción, mientras que descartan como herejía cualquier otra teoría que niegue alguna de dichas premisas fundamentales. La economía neoclásica tolera los resultados econométricos de cualquier clase, admitiendo así la validez estadística de relaciones que no puede incorporar a su armazón teórica, pero sin aceptar que de dichos resultados se deduzcan consecuencias contrarias a sus premisas fundamentales de conducta racional y de optimización.

Curiosamente, la postura neopositivista proporciona una justificación metodológica a estas teorías de tipo escolástico porque denuncia la explicación causal como ilusión metafísica y, en cambio, acepta las relaciones funcionales como resultado de una descripción de tipo fenomenológico. Así permite presentar los modelos económicos como simples parábolas, más o menos distantes de la realidad, pero que, tal vez, arrojen luz sobre otros hechos disímiles que sí podemos observar en la realidad (Samuelson, 1962), o justificar cualquier fórmula econométrica por su capacidad de pronóstico (Friedman, 1953), sin exigir que los supuestos sean realistas ni que tengan valor explicativo para otras situaciones similares.

Ambas propuestas metodológicas son inconsistentes con el propósito general de cualquier teoría: relacionar entre sí piezas explicativas de eventos concretos y reales sin recurrir a hipótesis *ad hoc* u otras formas de intervención divina. Y aunque en una teoría tienen cabida variables inobservables, no es lícito invocarlas *ad hoc*, pues solo son aceptables cuando sirven de puente con otras relaciones y variables verificables de esa misma teoría.

Es igualmente extraño que una teoría económica, en un campo donde todo es resultado de acciones y decisiones humanas, omita el estudio de dichas decisiones y de sus consecuencias, y se refugie en cambio en unos principios universales de racionalidad y optimización que parecen explicar lo que sucede como consecuencia impersonal e inevitable de unas circunstancias dadas.

Traza así el bosquejo de un mundo ideal, siempre en equilibrio, donde los individuos escogen con preferencias innatas que no modifican por experiencia propia ni por imitación ni emulación de sus semejantes, y terminan sustituidos por un *agente representativo*, como si todos fuesen iguales; y donde hay una multitud de empresas que dejaron de competir porque son óptimas e iguales a las demás. Colo-

cado todo esto en un entorno estático que solo puede cambiar por un progreso técnico exógeno que, cuando sobreviene, impacta a todas las empresas y a todos los trabajadores por igual, de modo que conserva el equilibrio de esa perfecta ausencia de competencia.

Ni siquiera discute la inconsistencia entre unas preferencias dadas y la posibilidad de innovaciones de producto que romperían cualquier esquema de elección prefijado, a menos que creamos que toda novedad ya fue prevista y "descontada" de todas las decisiones pasadas. De igual manera, pasa por alto la inconsistencia entre el supuesto de unas funciones de producción que guían la decisión empresarial y la realidad cotidiana de innovaciones de proceso y de conocimiento que alteran las condiciones de producción de mil maneras imposibles de prever, gestando rentas extraordinarias para algunos y la quiebra para otros. Así, al mismo tiempo que invoca las innovaciones como explicación del crecimiento y del desarrollo, niega sus consecuencias más evidentes.

En resumen, la teoría económica hoy dominante es un entramado lógico-deductivo, que parte de unos pocos axiomas indiscutidos –como su idea de racionalidad, su búsqueda de equilibrios y su principio de maximización–, descartando cualquier hecho sea incompatible con esos axiomas y adoptando cualquier epiciclo adicional que sea necesario para apuntalar su tolemaico esquema teórico-deductivo.

Una consecuencia directa del requisito metodológico de optimización dinámica es que sus modelos macroeconómicos solo pueden representar una clase muy especial de sistemas, los llamados hamiltonianos, que no admiten atractores de ninguna clase y que, una vez excluidas las soluciones periódicas, solo pueden tener puntos de silla. El ideal walrasiano es un equilibrio estable que se recupera por sí mismo después de sufrir una perturbación exógena. En cambio, la solución del óptimo dinámico no admite estabilidad alguna: dejado a sí mismo, este tipo de sistemas hamiltonianos profundiza cualquier perturbación exógena, hasta llegar a su autodestrucción.

Sus espacios de estados contienen algunos puntos especiales, los vértices de las sillas, y unas trayectorias privilegiadas que conducen hacia dichos vértices. Son las únicas combinaciones de estados que no llevan a la autodestrucción: si el sistema estuviera en uno de esos puntos estacionarios y no sufriera ninguna influencia externa, podría permanecer en él; igualmente, si se encontrara sobre una de las trayectorias privilegiadas y no sufriera perturbaciones exógenas, podría continuar su marcha hacia el punto estacionario, aunque a una

velocidad que tiende a cero a medida que se aproxima a él (es decir, que lo alcanzaría en un tiempo infinito).

Pero si comparamos los infinitos puntos del espacio de estados con esas zonas especiales de volumen nulo, podemos concluir que la probabilidad de estar en esos puntos, o en esas trayectorias, es cero. En ciencias naturales no se estudian estos sistemas porque rara vez sobreviven el tiempo suficiente para que podamos observarlos, y en ingeniería interesan solo porque es necesario entender cuándo podrían aparecer para evitarlos.

En cambio, la teoría neoclásica prefiere suponer la existencia de unas "expectativas racionales" que implican capacidades de información y de análisis infinitas para todos y cada uno de los agentes involucrados, de modo que pueden identificar esas trayectorias privilegiadas de probabilidad cero y situarse sobre ellas, compensando cualquier tipo de choque exógeno. Lo cual exige que cada individuo tenga un conocimiento exhaustivo de todo el sistema económico, así como una capacidad de cálculo infinita y una previsión perfecta, no solo de aquello que le afecta directamente, sino de lo que harán los demás agentes del sistema económico, y aún falta añadir las facultades necesarias para introducir los ajustes del caso en forma inmediata, antes de que los efectos desestabilizadores de la perturbación comiencen a empeorar el estado del sistema.

Por todas esas razones, las expectativas racionales constituyen un supuesto *ad hoc*, imaginado para salvar el principio universal de optimización que, aplicado al sistema dinámico, genera inestabilidad en vez de conducir al ideal deseado de equilibrio walrasiano.

Pedir algo de realismo en los supuestos llevaría a reconocer que todos actuamos con información limitada y bajo la incertidumbre de un futuro impredecible, en un mundo pleno de desequilibrios y en continuo proceso de cambio, condenados a corregir nuestros errores apenas percibimos las consecuencias indeseadas de decisiones previas.

El problema señalado no radica en la falta de evidencia estadística, y no se resolvería con más datos, reunidos bajo la guía de esa visión teórica neoclásica e interpretados luego como manifestación de supuestos óptimos y equilibrios.

La raíz del asunto está en el tipo de explicación que se considera válida: debemos escoger entre la racionalidad lógico-deductiva de esta corriente teórica y la falibilidad perfectible de la experiencia cotidiana, donde los seres humanos deciden y actúan, unos con los resultados esperados mientras que otros deben enfrentar las consecuencias de sus

errores y buscar remedios en sucesivas decisiones, todas igualmente sujetas al riesgo de error.

CAUSALIDAD EN ECONOMÍA

El conocimiento no consiste en una colección de hechos, sino que reside en las conexiones que concebimos entre ellos o, mejor, como no nos llegan hechos sino percepciones acerca de ellos, consiste en las conexiones entre perceptos. La importancia de esta distinción radica en que solo percibimos desde nuestras creencias, que guían a dónde mirar y determinan qué nos importa ver.

Las ideas aparecen integradas en un sistema explícito y, al menos en parte, conexo. Las creencias, en cambio, pueden ser inconexas, independientes unas de otras y, a veces, contradictorias; rara vez nos damos cuenta de ellas, y con mucha frecuencia no las conocemos; son verdaderos prejuicios que determinan cómo y qué vemos. Desde luego, las ideas conscientes pueden generar nuevas creencias y sustituir a las antiguas, pero este proceso es incierto y tortuoso, porque las creencias que tengamos en un momento dado guían la búsqueda de nuevas ideas y pueden sesgarlas. En pequeña escala, es un proceso análogo al cambio de paradigmas científicos (Kuhn, 1962).

Por otra parte, las conexiones que buscamos no están en los hechos, sino en nuestra mente; son el andamiaje de la conciencia y son imaginaciones, que no descubrimos sino creamos. Cada conexión postula una teoría acerca del mundo, pero como las teorías no se destilan de los hechos, cada hecho o conjunto de hechos admite varias conexiones imaginadas, es decir, varias interpretaciones teóricas.

La causalidad es un tipo de conexión, quizá el más simple porque se inspira en una experiencia corporal directa. Las dificultades de este concepto surgen de dos supuestos que es frecuente añadirle:

1. La unicidad, que construye secuencias o cadenas causales simples, aisladas, y

2. La linealidad, que permite el análisis en causas independientes que tienen efectos también independientes, lo que permite reconstruir el impacto conjunto de tales causas mediante la simple suma o superposición de sus efectos.

Otro problema, que parece difícil, pero resulta trivial, es que podemos citar muchas causas diferentes de un mismo fenómeno. Por ejemplo, cabe decir que la luz se debe a que alguien pulsó un interruptor en la pared de la habitación; o podemos extendernos y hablar de redes de transmisión eléctrica, de generadores y represas, y

añadir detalles sobre la instalación de las turbinas que impulsan a esos generadores y la extrema precisión de sus piezas; o podemos decir que cuando producimos una diferencia de potencial eléctrico aparece una corriente de electrones que, a su vez, genera una agitación térmica a nivel atómico que eleva la temperatura de un alambre delgado hasta el punto de incandescencia. Las tres causas son correctas: la primera puede ser suficiente en una disputa legal de responsabilidades; la segunda es pertinente para explicar las dificultades con que puede tropezar el desarrollo de un país, y la tercera es la que interesa al ingeniero que diseña fuentes luminosas: el contexto determina cuál es la explicación pertinente y cuál de las tres debemos considerar como causa y como parte de una explicación teórica que convenga a cada caso.

Más espinoso es que no existe una definición generalmente aceptada en las ciencias, ni un acuerdo filosófico sobre la causalidad. Aun en la versión que parece más usual en ciencias naturales, podemos hallar positivistas y neopositivistas, como Russell (1917), que la niegan por completo y afirman que la ciencia se basa en relaciones legales entre variables que se expresan mediante funciones o ecuaciones matemáticas. Otros filósofos adoptan una postura de parcial aceptación, insistiendo en su carácter de linealidad (Bunge, 2009), pero apenas como un caso especial entre las relaciones legales o funcionales que expresan alguna forma de determinismo, sea unívoco o probabilista. En otro extremo encontramos filósofos del conocimiento que defienden la causalidad como una propiedad específica, que amplía y complementa la simple relación lógica o funcional que pueda existir entre variables o eventos (von-Wright, 1971).

Aquí desarrollamos esta última posición, afirmando que la causa es una hipótesis que añadimos a la relación funcional para construir teorías: como veremos más adelante, podemos partir incluso de simples identidades sin valor explicativo, añadir supuestos de tipo causal y obtener entonces un enunciado verificable que es posible incorporar a una teoría explicativa más general.

En resumen, la hipótesis causal hace la diferencia entre la simple descripción fenomenológica y la teoría que pretende explicar los hechos observados: la primera proporciona una "caja negra" que, dados unos datos de entrada, entrega otros datos de salida sin que sepamos cómo ni por qué sucede esa transformación de datos en datos, como ocurre con los modelos ARIMA de pronóstico[1].

[1] Existe una conexión entre las series de Fourier, creadas para aproximar cualquier función o serie de datos con una precisión arbitraria, con solo añadir más términos, y los modelos autorregresivos. La transformación discreta

En cambio, la causalidad afirma que un evento genera o produce al otro, enunciado que va más allá de una simple asociación o de una correlación estadística, y que permite integrar la relación funcional dentro de un esquema más general y de tipo explicativo, es decir, en una teoría que podría señalar límites de validez para esa relación funcional, o sugerir nuevas variables y relaciones que enriquezcan la explicación o, en ciertos casos, descubrir alguna inconsistencia que obligue a revisar todos los supuestos fundamentales.

Este género de hipótesis causal hereda dos propiedades del origen sensorio-motriz del concepto de causalidad: la primera es la asimetría en el tiempo, que sitúa la causa antes del efecto o, a lo sumo, sincrónica con este; la segunda, igualmente esencial, es la asimetría que resalta el carácter opcional de la causa frente al carácter necesario del resultado y que consiente asimilar causas con decisiones. Estas dos asimetrías permiten construir cadenas causales, donde los efectos de una causa anterior se convierten en causas de efectos posteriores.

La asimetría entre causa y efecto admite la presencia de una reacción que en ciertos casos permite hablar de causación recíproca: no como una identidad, que reduciría la interacción a una relación funcional pura o a un simple asunto de definiciones, sino como una modificación del entorno que acompaña al efecto y puede alterar la causa inicial (porque no existen sistemas perfectamente cerrados y aislados, ni siquiera en los experimentos mejor diseñados), o bien como otra relación causal en sentido opuesto desencadenada por el efecto, que debe ser asimétrica en intensidad y, en general, posterior en el tiempo.

de Fourier muestra que se trata de problemas estrictamente duales el uno del otro y, por ende, equivalentes. La construcción de dichas series y transformadas de Fourier pone en evidencia que, al aplicarlas a un segmento finito de datos, en realidad estamos suponiendo que ese segmento se repite infinitas veces, mientras que el modelo autorregresivo alimenta la ilusión de que refleja el único segmento que, en efecto, fue observado. Así, la transformada deja claro que no permite predecir nada, porque presupone la repetición exacta, mientras que el modelo crea la ilusión de que sirve para predecir el futuro. Pero como la historia nunca repite el pasado, el pronóstico econométrico falla indefectiblemente, no porque el método de cálculo sea erróneo, sino porque la forma en que se anuncia –como representación de un proceso estocástico– crea una ilusión engañosa. Otros modelos más complejos de series de tiempo tienen también su equivalente en métodos de aproximación de funciones o de tratamiento de señales periódicas (Candy, 1988), con consecuencias similares, pero este sería tema de otro artículo, así como la discusión sobre las consecuencias de la linealidad de la mayoría de los modelos econométricos, o sobre las diferencias entre correlaciones estadísticas y conexiones de tipo causal.

La forma usual de causación recíproca es la retroalimentación en una cadena causal que se cierra sobre sí misma, lo cual hace posible que ciertos resultados actúen como reguladores de la causa inicial, a modo de mecanismo cibernético o de control automático: esto solo tiene sentido si existe una asimetría entre causas y efectos[2].

La mayor dificultad surge de que, en la práctica, las conexiones entre eventos ocurren en una red compleja con múltiples caminos de causalidad recíproca o retroalimentación. La interacción entre estos lazos causales, unos cortos y otros largos, algunos positivos o de refuerzo, y otros negativos o de atenuación, dificulta la comprensión de lo que puede estar sucediendo, porque no hay un hilo conductor que podamos seguir para explicar lo que sucede, sino varios hilos alternos que operan simultáneamente.

A veces podemos aislar una cadena corta, o una cadena dominante, o descubrir un comportamiento más o menos lineal, pero en el caso más general, el de mayor interés en economía, no es posible aislar cadenas causales independientes y, así como no podemos decidir cuál es el primer punto de un círculo, tampoco es viable una explicación ordenada a manera de encadenamiento secuencial: en cada eslabón de esa red causal subsiste la asimetría temporal, pero no es posible aislar causas independientes.

Este problema no ocurre únicamente en economía. La física comenzó con fenómenos que admitían unicidad (asimetría extrema) y linealidad; luego tropezó con fenómenos cuánticos que parecían aleatorios y solo recientemente admitió las consecuencias de la no linealidad y de una dinámica caótica en medio del determinismo.

La economía debe comenzar con redes causales, que no admiten análisis ni superposición lineal; además, su tema de estudio no tolera

[2] El concepto de reciprocidad ha sido invocado varias veces en la teoría económica y esto obliga a precisar algunas diferencias. El primero en hablar de "demandas recíprocas" parece haber sido Mill (1848), pero en un contexto estático, diseñado para justificar cierto equilibrio en el comercio externo. Más tarde, la misma expresión aparece en Young (1928), aunque esta vez como relación de tipo causal que forma parte de un proceso dinámico, semejante al que describió Veblen (1898) como causación cumulativa y que usó para explicar la transformación evolutiva de una economía. Descartamos aquí la expresión "demanda recíproca", así como otras semejantes, porque el uso de términos como demanda u oferta invoca significados y formulaciones nacidas de un enfoque estático, de equilibrio. Otro tanto sucede con el concepto de elasticidad que, en un proceso dinámico con innovaciones, solo se puede medir a posteriori y, por ende, no tiene valor explicativo ni predictivo. La idea de cadena causal cerrada que se presenta a continuación proviene de la teoría de sistemas y está libre de dichas connotaciones.

experimentos controlados. Tanto la econometría como las técnicas de análisis contrafactual (Spirtes et al., 2000; Eberhardt et al., 2010) pueden ayudar en ciertos casos, pero no podemos olvidar que llevan implícita la noción clásica de causalidad lineal.

La solución de estas dificultades consiste en cambiar de objetivo: el elemento básico de análisis de la red causal no es el eslabón, sino el proceso repetitivo (no periódico) que muestra una relativa estabilidad en ciertas variables porque uno o más lazos causales cerrados permiten la atenuación y eventual corrección de las desviaciones. La presencia de esa clase de estabilidad muestra que, por más compleja que sea la red causal, alguno de sus circuitos cerrados predomina sobre los demás y estabiliza esas variables en particular.

Se trata de una estabilidad dinámica, compatible con el movimiento y el cambio, incluso cuando estos ocurren a un ritmo exponencial. Para emplear un símil, en la mecánica de Newton la inercia se expresa como una estabilidad que deja en cero la segunda derivada del movimiento –la aceleración–, pero que es compatible con cualquier velocidad uniforme (primera derivada) y con cualquier ubicación en el espacio. En economía, la estabilidad se puede expresar entre tasas de crecimiento, es decir, entre velocidades de cambio, dejando libre de restricción el nivel o posición absoluta de la variable en cuestión, y es compatible con un crecimiento exponencial a tasa más o menos constante.

Esta forma de estabilidad dinámica, o de homeostasis como se dice en biología, sugiere la existencia de una restricción que debe respetar dicho sistema. Las identidades y las relaciones funcionales entre variables se cumplen en todo instante, mientras que las restricciones operan en forma indirecta, a través de una secuencia temporal de acciones y correcciones sucesivas, mediadas por una serie de variables intermedias que, en muchos casos, pueden pasar inadvertidas.

Si el periodo de observación es suficientemente largo, la restricción puede aparecer como una ecuación entre las variables principales que se cumple con un error de cierre pequeño, aunque nunca nulo porque el resto de la red causal también influye. A medida que acortamos el periodo de observación, ese error de cierre aumenta: primero, porque hay menos tiempo para que las fluctuaciones se amortigüen, y, luego, porque fases del ciclo causal comienzan a quedar incompletas.

En muchos casos de interés económico, la restricción empieza a verse en los datos trimestrales y alcanza una notable precisión en los datos anuales. Otras variables de ajuste más lento apenas dejan ver la restricción en los datos anuales y algunas, como las tasas de crecimiento de los grandes agregados, necesitan periodos de 10 o más años.

Si el indicio de restricción se refiere a una variable aditiva, es decir, tal que su valor total para n periodos sucesivos es la suma de los valores alcanzados en cada uno de los n periodos cortos, podremos reducir el error relativo de cierre con solo añadir más periodos. Esto se debe a que la restricción lleva implícito un proceso de corrección de errores que va eliminando los desajustes intermedios, así que, al cabo del periodo largo, el error final puede ser comparable, y aun menor, que muchos de los errores correspondientes a los periodos cortos intermedios.

Unas pocas fórmulas pueden aclarar esta afirmación mejor que muchas explicaciones verbales. Sean $a_i > 0$, $b_i > 0$ dos variables positivas y aditivas, diferentes entre sí, que deben cumplir la restricción de igualdad en el i-ésimo periodo corto, pero como hay perturbaciones debidas a otras variables y causas veremos pequeñas desviaciones:

$$a_i - b_i = e_i \neq 0 \tag{1}$$

La concatenación de varios periodos en un solo periodo largo daría agregados A y B, con un error de cierre $A - B = E$, y si no sabemos nada más de todos esos errores, solo podemos afirmar que $E \leq ne$, donde e es el mayor error absoluto hallado en los n periodos cortos. Como $B \cong nb_i$, el error relativo del periodo largo podría ser igual que el de un periodo corto, $E/B \cong e/b$, pero en muchos casos de importancia práctica es sensiblemente menor.

En primer lugar, si detectamos que los e_i tienen un componente sistemático, por ejemplo, una media $c \neq 0$, podemos redefinir la ecuación de la restricción así:

$$a_i - (b_i + c) = e_i \neq 0 \tag{2}$$

Los errores de cierre tendrán ahora media cero y signos diferentes, lo que hace una gran diferencia en el procedimiento de agregación porque se trata de errores accidentales y sus sumas convergen hacia una ley de Gauss o ley normal[3].

La suma de los errores de los n periodos ya no estará acotada por $E \leq ne$, sino por un límite bastante menor, del orden de $E \leq e\sqrt{n}$, así que el error relativo del periodo largo debe ser $E/B \cong (e/b)/\sqrt{n}$. Entonces, si la diferencia máxima entre a_i y b_i es del orden de un 10% de b_i, y agregamos 25 periodos sucesivos, la diferencia más probable entre las sumas $A = \Sigma\, a_i$ y $B = \Sigma\, b_i$ será del orden de un 2% de B.

[3] Aunque, en este caso bastaría aplicar el teorema de Chebyshev, que solo exige una cota máxima de error igual para todas las observaciones y que no necesita hipótesis alguna sobre distribuciones de probabilidad (Gnedenko, 1962).

Es una mejora notable, similar a la que se obtiene al sustituir una serie de tiempo por sus promedios móviles de 25 años: la nueva serie es mucho más suave porque gran parte de la fluctuación irregular desaparece. La diferencia está en que el promedio móvil es un filtro lineal que distorsiona todo el espectro de frecuencias de la serie original al combinarlas con las frecuencias propias de ese filtro, mientras que la lectura de una misma restricción en periodos de distinta duración no cambia ninguna de las relaciones funcionales que buscamos.

Pero hay otra clase de restricciones que permiten alcanzar una precisión mucho mayor en el agregado: si el error relativo en $(A - B)/B$ es igual o inferior al máximo error posible en el último periodo corto, $(a_n - b_n)/b_n$, desaparece el multiplicador $\sqrt{n}$ que teníamos en el cálculo de E porque ahora $E \cong e$.

Es lo que sucede, por ejemplo, cuando medimos la diferencia entre los préstamos recibidos y los pagos a la deuda realizados por una empresa o por un hogar, porque ambos tienen una capacidad máxima de endeudamiento que limita dicha diferencia en todo instante. En un caso así, la acotación del error relativo para el periodo largo será $E/B \leq (e/b)/n$.

Esta vez, de un error relativo máximo del 10% en cualquier período corto, conseguiríamos bajar a un módico 0,4% para el agregado B de 25 periodos.

Si hemos identificado una restricción como las descritas (y más adelante veremos algunos ejemplos), podemos simplificar el problema despreciando transitoriamente el resto de variables y relaciones de la red causal. Se obtiene así un modelo aproximado del sistema, que contiene unas pocas variables relacionadas directamente con la restricción, junto con algunos parámetros supuestamente constantes. El siguiente paso consiste en flexibilizar las relaciones funcionales así obtenidas, y caben dos estrategias:

1. Identificar otra restricción que comparta alguna de sus variables con la primera: en ese caso, el modelo comienza a combinar cadenas causales en una red compleja, o bien,

2. Convertir en nuevas variables algunos de los parámetros que aparecen como constantes en el modelo simple, buscando otras restricciones que deban cumplir esas nuevas variables.

En ambos casos, el nuevo modelo teórico tendrá más grados de libertad que el anterior y podrá generar dinámicas que antes no era posible ver. Pero es un modelo diferente y es preciso comprobar si todavía admite una estabilidad igual a la que caracterizaba al modelo anterior. En caso afirmativo, el nuevo modelo es una mejora del

anterior, pero si la respuesta es negativa, es preciso dar marcha atrás y revisar los cambios.

Esta precaución puede parecer extraña desde la perspectiva del análisis convencional, basado en una supuesta linealidad que garantiza la superposición por simple suma de los efectos a medida que añadimos causas independientes. Pero lo que sucede al combinar o flexibilizar procesos es que, si el antiguo modelo representaba cierto sistema, el nuevo modelo representa otro sistema diferente, más complejo y quizá no lineal, que puede comportarse de manera muy distinta del sistema simple.

Por ello, si el proceso anterior podía reproducir cierta forma de estabilidad observada en el mundo real, por ejemplo, una tasa de crecimiento estable con proporciones fijas entre ciertas variables clave, debemos comprobar si el nuevo proceso es compatible con esa misma forma de estabilidad, a pesar de que ahora tenga un mayor número de grados de libertad. Solo así, el modelo teórico se enriquece gradualmente sin perder las propiedades que sirvieron para validar las etapas anteriores.

HACIA UN NUEVO MARCO EXPLICATIVO

La economía real está siempre en desequilibrio, sujeta a innovaciones impredecibles; no persigue estados finales o normales, sino que traza su camino a medida que lo recorre. Construir una teoría que sirva de guía para diseñar políticas económicas efectivas parece un problema insoluble si partimos del comportamiento individual y pretendemos construir agregados aplicando el concepto tradicional de causalidad lineal, con sus métodos convencionales de análisis y posterior superposición de efectos independientes. Bajo esas condiciones, la única manera de simplificar la enorme variedad de comportamientos discordantes entre sí consiste en suponer que todos los individuos son iguales y crear la ficción de un "agente representativo", pero esa simplificación suprime las interacciones y elimina todas las conexiones de las que surgen las propiedades emergentes del sistema macro.

La alternativa propuesta, en cambio, apunta directamente a alguna de esas propiedades emergentes, la que queda en evidencia cuando hallamos una restricción macro que se cumple en forma aproximada. Luego hay que explicar su naturaleza y operación recurriendo a subagregados del mismo sistema macro, a las relaciones que describen sus comportamientos y a la estructura de relaciones que mantienen entre ellos, de donde nacen la capacidad autorreguladora del sistema

y, en muchos casos, las propiedades emergentes que solo existen en el nivel agregado.

A veces las relaciones en cuestión son de índole técnica, pero también pueden tener su origen en convenciones, instituciones sociales o normas legales vigentes.

En algunas ocasiones, podemos detener el estudio en el nivel de esos sub-agregados inmediatos, y aceptar un análisis puramente descriptivo de su comportamiento, a modo de "caja negra" que entrega ciertos resultados cuando recibe ciertas entradas. En general, es preferible considerar el primer nivel de sub-agregados como nuevos sistemas con componentes y estructura propios, para buscar después nuevas restricciones que los caractericen o una explicación de su comportamiento como propiedades emergentes de ese segundo nivel.

La extensión de un artículo apenas permite mostrar las posibilidades de este método seleccionando algunos casos relativamente simples, aunque es suficiente para encontrar rápidamente resultados fundamentales y de interés práctico que, por cierto, son muy diferentes de los propuestos por las teorías convencionales y, especialmente, por las de inspiración neoclásica.

La restricción presupuestal

Empecemos nuestro análisis macroeconómico con un sistema aislado, es decir, una economía cerrada donde, para simplificar, dejamos de lado el gobierno, con sus impuestos y sus transferencias, pero aceptamos como sector especial el financiero, que incluye bancos y mercados de capitales.

La primera observación es que las empresas pueden financiar parte de sus inversiones con recursos propios generados en el mismo periodo, pero necesitan pedir crédito para realizar los proyectos de inversión más importantes.

En principio, podrían emitir acciones nuevas y esta es una opción clara cuando nace la empresa, pero menos probable cuando está en marcha porque las nuevas emisiones diluyen la rentabilidad de los antiguos accionistas, que preferirán la emisión de bonos u otras formas de deuda. Por otra parte, la empresa misma prefiere el crédito porque su costo es transitorio, mientras que la emisión de acciones crea un compromiso permanente de dividendos adicionales. Por ello, solo hace emisiones secundarias cuando es imprescindible ampliar las garantías de capital propio. En ese caso, da preferencia a sus socios antiguos, aunque gran parte de las acciones nuevas terminan en manos de fondos de inversión y otras empresas intermediarias del sector financiero.

En resumen, el crédito y los bonos suelen ser la principal fuente de recursos, pero ambos tipos de endeudamiento están limitados por la capacidad de pago, medida a juicio de los bancos que dan el préstamo o intermedian la colocación de esos bonos en el mercado de capitales.

Los proyectos empresariales pueden ser de largo plazo y el pago del correspondiente crédito se distribuye en periodos sucesivos, pero la empresa debe restablecer su capacidad de pago antes de que pueda recibir otro crédito adicional. Bien sea como efecto de los criterios que aplica el banco, o bien porque, efectivamente, debe pagar la deuda con parte de sus ingresos netos, el endeudamiento de la empresa permanecerá acotado de manera que las cuotas anuales no sobrepasen una fracción de su ingreso neto anual. Otra forma de describir esta dinámica es que toda empresa, tarde o temprano, pagará sus deudas con sus ingresos netos.

En vez de considerar el proceso de una empresa a lo largo de varios años, podemos hacer un corte transversal y ver lo que sucede con un grupo grande de empresas en un mismo año. Cualquier agregado de empresas incluirá algunas que acaban de pedir crédito, otras que están pagando cuotas y otras a punto de cancelar sus deudas: esta rotación es un ciclo continuo, de modo que la deuda de ese agregado es una fracción del total de sus ingresos corrientes mucho menor que la fracción que la banca tolera para cada empresa por separado.

Además, las empresas de ese agregado que no están iniciando proyectos, colocan sus excedentes de caja en depósitos y valores del sector financiero, y estos recursos pueden servir para atender las necesidades de crédito de las empresas que sí invierten.

En un año determinado puede existir una diferencia entre las captaciones y las colocaciones de recursos de las empresas, aunque solo sea porque un auge sincroniza muchas decisiones de inversión, mientras que las recesiones las aplazan. Pero cuando consideramos un periodo más largo, de 10 o más años, las sumas de créditos y de inversiones tienden a nivelarse para cada empresa por separado y, por consiguiente, eso mismo debe suceder con las sumas de todas las empresas. La situación se acerca cada vez más al caso de un proceso ergódico, porque el error relativo de las diferencias disminuye y tiende a cero.

Pero esto significa que, en el largo plazo, las empresas financian sus inversiones con recursos propios y que los flujos de capital que reciben de propietarios y demás hogares son ocasionales y transitorios.

Es una conclusión muy diferente del supuesto tradicional de que las empresas invierten lo que los hogares ahorran, porque divide el

circuito financiero en dos ámbitos relativamente independientes. Las empresas y sus inversiones alimentan un primer circuito, mientras que los hogares mantienen un segundo circuito, casi independiente, donde unos hogares depositan sus excedentes en el sector financiero y hacen posible que otros hogares reciban créditos para financiar bienes duraderos, como la vivienda u otras compras a crédito.

Si incluimos las compras de activos y las colocaciones de los hogares en el mercado de activos o en empresas del sector financiero distintas de los bancos, el resultado sigue siendo el mismo: tenemos dos circuitos que operan en paralelo, con transferencias ocasionales entre ellos. Y, en los dos casos, la razón es la misma: a largo plazo, empresas y hogares deben pagar sus deudas con sus ingresos.

Hay muchas evidencias que apoyan estas conclusiones. Currie señala que, año tras año, la suma de inversiones del conjunto de empresas de Estados Unidos ha sido casi idéntica a la suma de las utilidades retenidas más los fondos de depreciación que quedaron en manos de esas mismas empresas: hay pequeñas diferencias en todos los años, pero no son sistemáticas y el balance se mantiene durante varios decenios (Currie y Sandilands, 1997). Esta diferencia casi nula respalda la sospecha de un proceso ergódico donde los promedios de un corte transversal anual son casi idénticos a los promedios de series de tiempo.

Otros estudios que intentan medir las fuentes de financiamiento de las empresas en diferentes países encuentran dos resultados que parecen antagónicos. Cuando miden lo sucedido en el conjunto de empresas de un país, encuentran que la mayor parte de los recursos de inversión son autogenerados, con un pequeño componente de crédito y otra partida aún menor de transferencias de ahorro de los hogares que, en ciertas épocas, puede ser negativa debido a la recompra de acciones por las empresas mismas (Corbett y Jenkinson, 1997). En cambio, cuando los estudios, en vez de usar los datos consolidados en un agregado nacional, suman los datos brutos empresa por empresa, encuentran que la inmensa mayoría de los proyectos de inversión se financian con crédito y fuentes del mercado de capitales (Hackethal y Schmidt, 2003).

Los dos resultados parecen contradictorios, pero la paradoja se resuelve por el requisito de que cada empresa debe pagar la mayor parte de sus deudas para restablecer su capacidad de endeudamiento futuro y, para que sea sostenible a largo plazo, debe hacerlo con los recursos que ella misma genera.

La restricción de ventas

Veamos ahora otra restricción aún más simple que la anterior, tanto que parece una simpleza pueril, y es que, en la economía moderna, las empresas producen todo lo que pueden vender.

Desde luego, siempre mantienen una capacidad instalada ociosa para atender pedidos extraordinarios, aprovechar el mercado que deja uno de sus competidores por accidentes o huelgas, o, simplemente, para facilitar su crecimiento futuro. Además, mantienen inventarios de materia prima, bienes en proceso y terminados en prevención de dificultades de aprovisionamiento y de fluctuaciones menores o transitorias en sus ventas, inventarios que representan una fracción relativamente pequeña de su producto anual y que rotan continuamente. Pero ninguna empresa tiene por objetivo producir para acumular inventarios: cuando no vende, reajusta sus planes de producción a lo que puede vender.

Si consideramos el agregado de todas las empresas, las inversiones representan ventas y compras dentro de ese mismo sector que, además, son una necesidad derivada de las demás ventas, es decir, de las que las empresas hacen a los demás sectores.

El verdadero motor de la producción viene a ser, entonces, el volumen de ventas a los hogares y, en las economías modernas, la mayor parte corresponde a bienes de consumo masivo.

Los propietarios, es decir, quienes reciben ingresos de capital u otras fuentes similares, pueden tener una gran concentración de capital y de ingresos, pero su capacidad de consumo es limitada: lo que excede esa capacidad se traduce en colocaciones en el sector financiero. Por esa razón, los periodos de concentración creciente del ingreso coinciden con auges especulativos en bolsas de valores y bienes inmuebles, es decir, de activos preexistentes que no forman parte del producto corriente, así como tampoco suman al mismo las valorizaciones contables que puedan experimentar dichos activos.

Como consecuencia, las ventas de las empresas a los hogares dependen en pequeña medida de los consumos suntuarios de propietarios de capital, mientras que la mayor parte del producto corriente refleja el consumo masivo de los hogares de trabajadores, tanto asalariados como independientes. Todo esto se puede resumir diciendo que la economía moderna es una sociedad de consumo de masas.

Supongamos ahora que las innovaciones tecnológicas permiten aumentar la productividad del trabajo, es decir, obtener más producto físico con la misma cantidad de trabajo. Para que las empresas pue-

dan vender ese producto físico adicional es necesario que aumente el ingreso real de los trabajadores, así que, o bien las empresas bajan los precios de sus productos, o bien tienen que pagar salarios más altos. Pero esta condición significa, simplemente, que todo aumento de productividad debe trasladarse al ingreso real de los trabajadores para que sea sostenible en el largo plazo.

La economía tradicional, que centra su atención en las conveniencias de las empresas productoras, suele promover políticas de reducción del costo laboral con la idea de que un margen de ganancia mayor inducirá más inversiones y generará más empleo. Pero el análisis de circuito, basado en la restricción de que solo se produce lo que se puede vender, predice que las reducciones del costo laboral tendrán, como consecuencia inmediata, una disminución de las ventas y, por consiguiente, un aplazamiento o una cancelación de las inversiones: en vez de crecer más rápido, comenzará una fase recesiva que perjudicará a empresas y trabajadores por igual.

Desde luego, hay dos estrategias que podrían paliar las consecuencias negativas de recortar los ingresos del trabajo, al menos por un tiempo limitado. Una consiste en tolerar una concentración progresiva del ingreso y confiar en que aumente el consumo suntuario en cantidad suficiente para compensar la pérdida de capacidad adquisitiva del trabajo, pero la concentración impone un límite a esta salida porque el número de hogares de alto ingreso es relativamente pequeño: sería necesario regresar al lujo aristocrático del siglo XIX, con sus ejércitos de servidores privados y sus mecenazgos para que los más ricos puedan generar una capacidad alterna de compras masivas. La otra salida consiste en abrir la economía al comercio externo y aprovechar el bajo costo laboral para competir con otros países, aunque es muy probable que haya costos aún menores en otro lugar, lo que puede iniciar una competencia hacia el mutuo empobrecimiento nacional; por esta vía, los países pobres terminan con una población que no puede educarse ni mantener en sus productos el nivel de calidad exigido por los países ricos, cuyo mejor nivel de vida les permite importar, pero que acaban comerciando entre sí por razones de innovación y calidad.

La inversión y los fondos privados de pensiones

La expansión del sector financiero, que va de la mano con el crecimiento de las economías, tiene mucho que ver con la canalización del ahorro de los hogares, primero con las compañías de seguros, luego con una amplia variedad de fondos de inversión y, recientemente, con los

fondos privados de pensiones. No hay duda de que esta concentración de recursos facilita la financiación de grandes proyectos de inversión, pero las dos restricciones que acabamos de discutir ponen en duda su eficacia para aumentar el ritmo o el volumen de la inversión en el agregado de la economía.

La economía tradicional concibe la inversión como una consecuencia del ahorro de los hogares, que estos transfieren finalmente a las empresas; así, cualquier concentración de los ahorros y, aún mejor, cualquier ahorro obligatorio que se concentre en fondos privados de pensiones, debe acelerar la inversión y el crecimiento.

Pero la restricción que actúa entre producción y ventas indica que las inversiones dependen de las ventas: si estas aumentan, las empresas estarán dispuestas a invertir y buscarán las fuentes de financiamiento que necesiten para ese fin. En cambio, si las ventas flaquean, aplazarán sus inversiones porque, de otra manera, estarían produciendo para acumular inventarios y pérdidas.

Todo intento de aumentar la masa corriente de ahorros de los hogares redunda en una disminución equivalente de su gasto corriente, es decir, en un menor volumen de ventas, de modo que esos ahorros son compensados muy pronto por una caída de las ventas, seguida poco después por un descenso de la inversión, del empleo y del ingreso.

La restricción presupuestaria, que indica la presencia de un circuito de financiamiento entre el ahorro de unos hogares y el gasto en bienes durables de otros hogares, sugiere que, si excluimos la emisión primaria de medios de pago, todavía es posible reorientar el ahorro monetario de los hogares y concentrarlo en ciertos fondos financieros sin consecuencias negativas para el crecimiento, pero solo si se reduce el ahorro que antes fluía por otros canales para financiar la compra de vivienda, vehículos, electrodomésticos y demás bienes durables. En tal caso, no habría un aumento del ahorro, sino solo una redistribución del mismo.

Esto debilita considerablemente el argumento en favor de los fondos privados de pensiones como creadores de un ahorro que permitiría acelerar la inversión y el desarrollo.

Quedaría aún la consideración de que el régimen de prima media es insostenible porque exige aportes del presupuesto nacional a medida que crece el número de pensionados y disminuye la proporción de población en edad de trabajar. Pero es fácil ver que los fondos privados tropiezan con dos dificultades de importancia similar.

La acumulación de recursos y la capacidad del fondo privado de pensiones para financiar grandes proyectos de inversión –como obras

públicas u otras colocaciones seguras– es un asunto del horizonte de análisis: si proyectamos a 10 o 20 años, el fondo estará recibiendo cuotas de muchos afiliados y pagando pensiones a un número muy pequeño de beneficiarios, pero, como la transición demográfica continúa su marcha y cada vez hay una menor proporción de nacimientos, los pagos de pensiones alcanzarán el monto de los recaudos al cabo de 30 o 40 años, y poco después los superarán. De ahí en adelante, el fondo privado comenzará a desacumular activos y la presión de sus ventas empezará a tener un impacto negativo sobre los precios de dichos activos en el mercado de capitales, acelerando el deterioro de las finanzas de cada fondo.

Llegará un punto en que necesitarán una inyección de recursos del presupuesto nacional para sostener sus pagos de pensiones y evitar una caída de las ventas agregadas en toda la economía: se habrá cambiado un flujo gradual de transferencias de impuestos a pensiones en el régimen de prima media, por una cascada de transferencias masivas a cargo de generaciones futuras.

Otra reflexión surge de combinar las dos restricciones anteriores. La idea generalizada de que la pensión es un subsidio o un regalo al jubilado es un prejuicio "micro" que no tiene sustento desde la perspectiva social o del agregado: antes de que el individuo se retire por vejez, tiene un ingreso que contribuye a mantener el circuito de producción y ventas y, si en un mismo año se retiran unos cuantos miles de individuos sin pensión, sus gastos caerán drásticamente y disminuirán las ventas agregadas. Así será poco probable que las empresas sostengan el nivel anterior de empleo, es decir, que sustituyan a los jubilados por jóvenes.

Si esos jubilados están afiliados a un fondo privado, el impacto negativo será menor en el lado de las ventas, porque recibirán unos pagos contra sus fondos acumulados, pero el impacto negativo no desaparece porque, al no acumular capital y rendimientos como antes, el impacto sobreviene por el lado del financiamiento de las inversiones que puede hacer el fondo, en especial si ya se llegó a la fase de desacumulación de los fondos privados.

Lo que muestra este análisis es que las pensiones actúan como un subsidio de la sociedad a sí misma: en vez de ser una carga, son pagos que mantienen el nivel previo de actividad económica y, desde esa base, facilitan el crecimiento ulterior del empleo y del producto agregado. En cambio, eliminar, gravar o reducir las pensiones disminuye en forma automática el nivel de actividad económica existente, con un probable deterioro del empleo.

La brecha inflacionaria

Volvamos a la economía cerrada y dividamos el sector empresas en dos grupos: uno que innova regularmente y otro que busca protecciones para mantener sus formas de producción, es decir, sus instalaciones, las patentes en uso y las demás inversiones ya realizadas. Pueden ser empresas suficientemente grandes para hacer "lobby" o pueden ser gremios de productores medianos y pequeños; en ambos casos, el argumento suele ser el mismo: hay que proteger los empleos existentes con barreras o subsidios que preserven la tecnología en uso y las inversiones ya realizadas.

El subsector innovador puede aumentar su producción para aprovechar la ventaja tecnológica que ha conseguido y estaría dispuesto a bajar precios para facilitar ventas adicionales, si fuese necesario. Por otra parte, como la innovación exige trabajadores más calificados, empieza a ofrecer mejores salarios para atraerlos. Arranca entonces una competencia entre este sector y el tradicional, que no crece al mismo ritmo y que, si quiere retener a sus trabajadores, también debe aumentar los salarios pagados.

La estrategia del sector protegido será siempre buscar más protecciones para compensar sus mayores costos laborales y esto presiona al alza la mayoría de sus precios. Lo que tenemos aquí es una presión inflacionaria generada por el proteccionismo, tanto mayor cuanto más rápido crezca la productividad del subsector innovador.

Es posible que, más adelante, aparezca una espiral de precios y salarios que acelere la inflación, pero su origen no es la presión sindical, y no se resuelve con "políticas de ingresos y salarios" que solo empobrecen a la mayoría de los hogares y frenan el crecimiento de la economía.

Durante el siglo XIX, con poca intervención de los Estados y un sistema monetario frenado por el patrón oro, el traslado de los aumentos de productividad al salario real tuvo lugar mediante una reducción casi continua de los precios nominales y un aumento relativamente lento de los salarios nominales. Pero después de la Segunda Guerra Mundial, en una economía dominada por el temor a la deflación y con un sistema monetario más flexible, el ajuste de la capacidad adquisitiva para absorber los efectos del aumento de productividad ocurrió principalmente por una sucesión de espirales inflacionarias, donde los salarios nominales pudieron subir algo más rápido que los precios nominales, al menos mientras duraron las políticas de pleno empleo (Sylos-Labini, 1991a y 1991b).

Asimetrías en la capacidad de decisión

El análisis precedente combinó una restricción –que las empresas producen solo lo que pueden vender– y una asimetría causal que está implícita en cualquier decisión o acción. El punto clave es que algunos agentes están en condiciones de decidir, mientras que otros solo pueden reaccionar ante las consecuencias de tales decisiones.

En el ejemplo anterior, las empresas pueden escoger sus precios de venta, el número de empleados que contratan y el nivel de salarios que ofrecen. Los compradores reaccionan luego decidiendo cuánto adquirir a esos precios y, más tarde, las empresas verán si mantienen sus precios, lanzan campañas de rebajas o innovan con modelos más atractivos. Los trabajadores también tienen un papel pasivo porque solo pueden decidir si aceptan o no el empleo que les ofrecen.

Esta clase de asimetrías impide hablar de mercados donde se enfrentan una oferta y una demanda que ajustan precios hasta alcanzar un punto de equilibrio; también impide hablar de un "mercado de trabajo" donde empresas y trabajadores regatean hasta llegar a un acuerdo.

La dinámica del trabajo y de los salarios depende entonces del balance entre las necesidades de trabajadores que tengan las empresas y su disposición a competir una contra otra para atraer empleados de la calificación deseada, o incluso sustraerlos de otras empresas ofreciéndoles alguna ventaja salarial.

En consecuencia, los salarios suben en las fases de auge y bajan en las recesiones. Si hay poco desempleo en las calificaciones laborales que requieren las empresas, el aumento de salarios es rápido, pero esta situación es compatible con alto desempleo en otras calificaciones laborales porque no se pueden sustituir pocos calificados por muchos no calificados. Claro que, más adelante, cuando exploren nuevos proyectos de inversión, las empresas buscarán tecnologías que requieran menos trabajadores calificados para frenar así el aumento de los salarios, pero esto no suele ser suficiente para reducir el desempleo de los no calificados.

Existe una relación entre salario, distribución del ingreso, productividad del trabajo y desempleo, pero depende de la interacción entre varias restricciones y cadenas causales, y es demasiado compleja para analizarla en este artículo. El lector interesado puede consultar Lorente (2018).

CAUSALIDAD Y RELACIÓN FUNCIONAL

Cambiemos ahora de tema para mostrar la diferencia entre una simple identidad sin valor explicativo y una relación funcional que puede integrarse dentro de una teoría propiamente dicha.

Empecemos construyendo una identidad contable a partir de definiciones y mediciones que proyectan una relación micro familiar a otra relación entre agregados nacionales. En forma sumamente esquemática, es el mismo procedimiento que sirvió para diseñar las Cuentas Nacionales.

La identidad micro de partida es la cuenta de ingresos netos de una empresa que tiene un solo tipo de trabajador y un equipo de capital homogéneo, después de descontar lo que pagó por los insumos y equipos del periodo:

$$y = wl + rk \tag{3}$$

donde w es el salario, l el número de trabajadores y r la rentabilidad que mide sobre el capital contable k de que dispone. La contabilidad de cada empresa incluye las compras de insumos y equipos, pero podemos omitir esos detalles porque desaparecen en el agregado: cada una de estas compras es una venta de otras empresas y se anulan mutuamente al consolidar las cuentas del sector. Varios modelos macro de uso frecuente se limitan a usar la misma fórmula para el agregado, escrita así:

$$Y = wL + rK \tag{4}$$

confiando en las Cuentas Nacionales y en las estadísticas laborales para medir algunas de esas variables, y definiendo otras con procedimientos semejantes a los que usaría el empresario micro.

Las Cuentas proporcionan estadísticas del producto agregado Y y miden en forma más o menos directa[4] el ingreso de los trabajadores $[wL]$, de modo que basta conseguir el dato de empleo global L para deducir un salario medio $w = [wL]/L$. Por simple diferencia hallamos entonces el ingreso atribuible al capital:

$$[rK] = Y - [wL] \tag{5}$$

[4] Las Cuentas obtienen información de las nóminas pagadas por las empresas y calculan el ingreso de los hogares independientes, que llaman ingresos mixtos. Hay que separar luego este último en una parte atribuible al trabajo y otra atribuible al capital utilizado. Una aproximación usual consiste en calcular el salario medio de las nóminas y aplicarlo al número de trabajadores independientes para estimar el componente laboral. El resto del ingreso mixto se atribuye al capital.

pero no tenemos información directa del capital K. En su lugar, las Cuentas calculan la inversión bruta del periodo, llamada formación bruta de capital, FBK, que podemos usar para definir una serie de capital con un procedimiento similar al que aplica una empresa en su contabilidad. Debemos escoger un monto inicial, $K(0)$, para el primer periodo de la serie y una tasa de depreciación media d, con lo cual calculamos el capital K que habrá al comienzo del siguiente periodo:

$$K(t + 1) = (1 - d)K(t) + FBK(t) \quad (6)$$

Luego calculamos la rentabilidad media de este capital:

$$r = [rK]/K$$

Hasta este punto no hemos introducido ningún supuesto de teoría económica, sino apenas construido una identidad contable: es indudable que se cumple con total precisión, porque adoptamos las convenciones y definiciones que así lo garantizan.

De aquí en adelante es posible construir nuevas interpretaciones de las variables para transformar la identidad en una relación funcional e integrar esta relación dentro de una teoría.

La teoría convencional

La interpretación convencional se remonta más allá de Adam Smith, pero logra su forma completa con la defensa del empresario capitalista que hace David Ricardo, al incluir el capital como un factor de producción con igual nivel que el factor trabajo (Ricardo, 1821).

Podemos representarla como una relación causal de la forma:

$$\{K, L\} \rightarrow Y \quad (7)$$

pero hay distintas maneras de interpretar este simbolismo.

La escuela neoricardiana describe una tecnología lineal con coeficientes técnicos fijos que miden la cantidad de diferentes clases de capital físico $K_1, K_2,...$ y de trabajo $L_1, L_2,...$ necesarias para obtener una unidad de cada producto físico Y_j.

El conjunto de técnicas es un par de matrices: una de trabajo y la otra de insumos físicos. Las columnas de esta última suman menos de la unidad, de manera que la producción de cada bien es siempre mayor que el volumen consumido. Con un producto por cada técnica, es posible hallar precios de equilibrio para cada producto y, si se acepta una relación fija entre los salarios de los distintos tipos de trabajo, se encuentra una relación lineal sencilla entre el nivel general de salarios y la rentabilidad del capital.

De esta manera, la descripción respeta la heterogeneidad de bienes de capital y calificaciones de trabajo, pero, lo mismo que sucedió con el análisis de Ricardo, la simplicidad de las ecuaciones se pierde cuando hay máquinas que duran varios periodos sucesivos, o cuando existe producción conjunta de varios bienes a la vez. Además, la canasta-índice que podría servir como unidad de medida para comparar alternativas de distribución del producto agregado cambia con cualquier innovación de proceso o de producto.

La simplificación usual de esta idea es el modelo lineal *AK*, que reduce la explicación del producto a la disponibilidad de capital físico y deduce el empleo aplicando un coeficiente de proporcionalidad fijo, es decir, como una consecuencia del capital disponible en cada instante. Aunque parezca una hipótesis extrema, muchas otras teorías atribuyen una importancia excluyente a la acumulación de capital, incluso cuando se trata de explicar el desarrollo de los países, mientras que tratan el trabajo como un factor excedentario, siempre disponible.

Gran parte de los modelos keynesianos simplifican aún más y utilizan un capital homogéneo, representado por una variable unidimensional, del cual derivan el empleo y el producto usando coeficientes técnicos constantes. Esta práctica excluye los efectos del cambio técnico, aunque es frecuente hallar textos donde dichos modelos se aplican erróneamente al análisis de periodos largos, cuya duración es mucho mayor que el "corto plazo" de Marshall[5].

La corriente principal de la teoría económica conserva la idea de un producto determinado por la disponibilidad de capital y trabajo al postular la existencia de una función de producción agregada, continua y cóncava en las dos variables, que representa el máximo producto obtenible con cada combinación de factores K y L:

$$Y = F(K, L) \tag{8}$$

A este supuesto añade otro de comportamiento racional que aprovecha toda la disponibilidad de cada factor, de modo que cualquier aumento de K o de L se traduce de inmediato en un reajuste correspondiente del producto Y, y así transforma la relación funcional en una explicación del producto de tipo causal, más estricta aún que las relaciones lineales con coeficientes fijos de las teorías antes mencionadas que, en ciertos casos, podrían dejar sobrantes ocasionales de uno u otro factor.

[5] El corto plazo de Marshall no es un intervalo determinado de tiempo físico, sino un "tiempo virtual" que él define como un lapso insuficiente para cambiar equipos y técnicas. Así evita discutir las dificultades del cambio técnico y de los rendimientos crecientes (Marshall, 1920). Keynes adopta tácitamente la misma definición de corto plazo en su *Teoría general*.

La interpretación de las variables *Y*, *K* y *L* como vectores compuestos por distintos bienes y tipos de trabajo, así como la heterogeneidad de las técnicas en uso, fueron objeto de un análisis minucioso en los años 60 y fuente de la discusión entre "los dos Cambridge" que puso en evidencia las dificultades del análisis neoclásico de agregados, como se puede ver en Harcourt (1972). Por la misma época, el examen del problema de agregación de funciones de producción micro para encontrar una función macro halló condiciones estrictas que equivalían a suponer que cada empresa debía utilizar un vector igual de tipos de trabajo y de bienes de capital para obtener el mismo vector de productos, usando funciones de producción micro que solo podían diferir en un coeficiente multiplicativo, es decir, proporcionales entre sí (Fisher, 1969). Esto equivale a tener un producto único, obtenido con un trabajo uniforme y un solo bien de capital, es decir, admitir que *Y*, *K* y *L* representan variables unidimensionales y no vectores de elementos heterogéneos.

Otro supuesto clave es el de competencia perfecta, entendida como la presencia de muchas empresas que deciden su producción sin restricciones de acceso a tecnología ni a capital. Unido al supuesto de conducta optimizadora y a la homogeneidad de producto y de capital, esto significa que todas las empresas deben escoger la misma técnica y el mismo volumen óptimo de producción, es decir, que todas son idénticas entre sí.

De este modo, las dificultades de agregación y el supuesto de competencia perfecta convierten al agente representativo en un supuesto necesario, en vez de ser una simplificación transitoria. Y como todas las empresas deben ser iguales, la única manera de aumentar el producto consiste en añadir más empresas idénticas, así que toda función de producción debe ser homogénea de primer grado.

La resistencia a abandonar la competencia perfecta se debe a que es un requisito para que los precios de mercado estén dados por las derivadas parciales de la función de producción con respecto al factor respectivo. Pero esto último trae una consecuencia indeseable: la homogeneidad de primer grado de la función *F* determina que la remuneración de los factores *K* y *L* agota el producto *Y*, y así es imposible destinar una parte de este para financiar actividades de investigación y desarrollo, tanto de procesos como de productos.

Este es el modelo neoclásico de crecimiento (Solow, 1956), que resuelve el problema de la distribución del producto como una consecuencia de la tecnología, pues la forma de la función de producción, combinada con el supuesto de competencia perfecta, determina los precios así como el volumen óptimo de producto.

Sin competencia perfecta, la tecnología no podría determinar los precios, pero cuando aceptamos ese requisito todo el cambio técnico debe ser exógeno, o provisto por el Gobierno con fondos públicos, porque, como ya dijimos, el modelo no deja ningún excedente de producto que las empresas del sector privado puedan utilizar para financiar investigación ni desarrollo de procesos o productos nuevos.

Esto llevará más tarde a introducir los modelos de crecimiento endógeno que eliminan el supuesto de competencia perfecta y remuneran la investigación a través de un sistema de patentes que crea monopolios transitorios. Pero esta solución también tiene dificultades porque la tasa de crecimiento depende entonces del número de investigadores contratados y, si ese número aumenta a ritmo exponencial, como sucedió durante todo el siglo XX, el modelo predice un crecimiento a escala, es decir, doblemente exponencial, en contradicción con la evidencia estadística disponible.

Sin heterogeneidad de empresas tampoco hay diferencia entre el análisis micro y el macroeconómico, es decir, no es posible una micro-fundamentación de los modelos agregados, porque la única diferencia entre esos dos niveles es un factor de proporcionalidad. Por esa misma razón, no es posible hallar una explicación consistente de las propiedades estrictamente macro, porque no existe ninguna interacción de agentes heterogéneos que pueda dar lugar a propiedades emergentes del sistema.

La única defensa que le queda a la idea de función de producción agregada es que existe alguna evidencia econométrica en su favor. Precisamente, el modelo de Solow es bastante popular en las teorías del crecimiento porque, con solo añadir un factor exponencial que "represente" el efecto del cambio técnico exógeno, permite obtener un buen ajuste en muchos países, para periodos relativamente largos, usando una función de producción Cobb-Douglas, que es homogénea de primer grado.

El modelo de Solow consta de tres ecuaciones: una que describe el crecimiento demográfico, $L = L_0 e^{nt}$; la segunda es la regla de cálculo del capital K donde se supone que la inversión bruta es proporcional al producto, $FBK = sY$, y la tercera es la función de producción, $Y = ae^{bt}K^{c}L^{1-c}$, donde b representa la tasa de cambio técnico exógeno que beneficia por igual a todas las empresas existentes.

Con el supuesto de competencia perfecta se deducen los precios mediante derivadas parciales $r = \partial F/\partial K$, $w = \partial F/\partial L$ y se muestra que, si la economía parte del punto de equilibrio (donde la proporción K/L

es constante), el crecimiento subsiguiente es balanceado (con K/Y constante) y exponencial a la tasa $g = b + n$.

También se deduce que la participación del capital en el producto, rK/Y, es igual al exponente de K en la función de producción Cobb-Douglas; así, la participación del trabajo, que recibe el remanente del producto, es $wL/Y = 1 - c$, lo que demostraría que la distribución del producto es una consecuencia de la tecnología en uso.

Desde luego, la evidencia econométrica no es totalmente favorable porque, si aceptamos la información de Cuentas Nacionales, el parámetro c debe estar cerca de 1/3, tanto en los países desarrollados como en los subdesarrollados, mientras que muchos ajustes, en especial los de países en desarrollo y los que usan datos de panel que mezclan un buen número de países, dicen que c debe estar cerca de 2/3, es decir, una distribución del ingreso exactamente opuesta a la medida en forma directa.

Las predicciones del modelo son muy sensibles a estos valores. Si $c = 2/3$, las comparaciones entre países ricos y pobres indicarían diferencias razonables en la dotación de capital y en su rentabilidad; por ejemplo, un país rico que tenga 10 veces el ingreso per cápita de uno pobre debe tener un capital por trabajador unas 32 veces mayor, mientras que la rentabilidad de dicho capital debe ser unas 3,2 veces mayor en el país pobre que en el rico. Pero si usamos $c = 1/3$, las diferencias son abismales; el país rico debe tener 1.000 veces más capital por trabajador que el país pobre, mientras que la rentabilidad debe ser 100 veces mayor en el país pobre. Aunque se intente subsanar parte de estas diferencias con argumentos de eficiencia relativa o de restricciones al conocimiento técnico, es inevitable concluir que el capital debería trasladarse masivamente a los países subdesarrollados, acabando en pocos años con cualquier diferencia de ingreso frente a los desarrollados; en cambio, la evidencia apunta a transferencias en sentido opuesto. De igual manera, los países pobres deberían atraer capital humano de los países ricos (Lucas, 1990), porque su precio sería mucho menor donde es abundante, pero también en este caso las migraciones suelen ir en sentido contrario.

Desde hace años (Fisher, 1971), se sospecha que el éxito del ajuste econométrico se debe a razones distintas de las que aduce el argumento neoclásico. Podemos adaptar un argumento sencillo de (Felipe y McCombie, 2007) para mostrar que existe una relación funcional parecida a la función de producción neoclásica, pero mucho más precisa y que no necesita ningún supuesto de teoría neoclásica.

La identidad contable con la que empezamos esta sección:

$$Y = rK + wL \tag{9}$$

se debe cumplir tanto al inicio como al final de cualquier periodo, aunque haya habido crecimiento o cualquier tipo de reajuste. Si suponemos que esos cambios son pequeños en comparación con el valor de la variable respectiva, podemos escribir la relación diferencial:

$$dY = Kdr + rdK + Ldw + wdL \tag{10}$$

que no es exacta, pero sí una aproximación de primer orden, con excelente precisión si, como sucede en la práctica, se trata de cambios anuales del 2 al 5 por ciento.

Dividiendo por el producto Y y reagrupando términos, podemos obtener su equivalente en tasas de cambio:

$$dY/Y = (rK/Y)(dr/r) + (rK/Y)(dK/K) + (wL/Y)(dw/w) + (wL/Y)(dL/L) \tag{11}$$

Ahora bien, la restricción de producción y ventas sugiere que, a pesar de las innovaciones técnicas y demás cambios graduales, las participaciones del capital y del trabajo en el producto serán estables mientras no sobrevengan cambios importantes en la concentración del ingreso. Por ello, podemos esperar que, en muchos países y en periodos bastante largos, las participaciones observadas fluctúen ligeramente alrededor de una constante y, para simplificar, vamos a suponer que c permanece constante durante ese tiempo.

La aproximación anterior se escribe entonces como:

$$dY/Y = cdr/r + cdK/K + (1 - c)dw/w + (1 - c)dL/L \tag{12}$$

Pero esta es una ecuación diferencial en variables separadas que podemos integrar de inmediato para obtener:

$$Y = hr^c w^{1-c} K^c L^{1-c} \tag{13}$$

expresión que difiere de la función de producción Cobb-Douglas antes citada en el cambio de la exponencial exógena, ae^{bt}, por el producto $hr^c w^{1-c}$. El ajuste econométrico de esta fórmula logra una precisión mucho mayor que la función de producción de Solow, porque no es más que una aproximación de primer orden a la identidad contable original.

Avancemos un poco en el análisis de la nueva expresión $hr^c w^{1-c}$. Sabemos que debe crecer exponencialmente y que la rentabilidad del capital, r, varía poco de un año a otro: desde luego, nada parecido a un aumento exponencial sostenido en el tiempo.

Por consiguiente, la variable que debe crecer a tasa exponencial es el salario real medio w y, efectivamente, es la misma conclusión a

que nos llevó la restricción de producción y ventas al comienzo de este artículo: la remuneración del trabajo debe crecer al mismo ritmo que el producto para que la moderna sociedad de consumo de masas sea sostenible.

Es tentador decir que el modelo neoclásico está sesgado hacia el punto de vista de la oferta y que bastaría reinterpretar la productividad total de factores como el término que falta del lado de la demanda para, así, restablecer el correcto balance entre los dos puntos de vista. De paso, estaríamos descifrando la misteriosa "productividad total de factores", que la teoría neoclásica jamás ha podido explicar; pero ninguna de estas conclusiones es razonable ni, menos aún, lógicamente necesaria.

La nueva fórmula del producto *Y* sigue siendo la expresión de una identidad, válida por definición, aunque solo aproximada por construcción; así que se trata de una relación funcional sin significado propio, que admite múltiples interpretaciones.

En particular, no tiene sentido insistir en la idea de funciones de producción porque la simple disponibilidad de factores no es garantía de que se decida usarlos, como lo prueba cualquier depresión económica.

Podría existir una relación causal entre las variables que aparecen en la identidad, pero es mucho más probable que algunas, o todas, sean efecto de otras variables que no aparecen en dicha fórmula. Por ejemplo, la búsqueda de una teoría mejor puede comenzar con la observación, antes comentada, de que la decisión de producir depende de que sea posible vender ese producto.

UNA INTERPRETACIÓN ALTERNA

No es posible exponer aquí una teoría alterna completa, pero sí trazar parte del esquema causal sugerido por las restricciones antes explicadas y por las críticas a la interpretación convencional de los factores que acabamos de ver[6].

Como decíamos, el punto de partida es la posibilidad de venta, real o imaginada por el empresario, porque ese cálculo determina su decisión de invertir.

Las inversiones son verdaderos paquetes tecnológicos que combinan ciertos equipos y ciertos tipos de trabajo de diferentes calificaciones, así que tanto el capital contable como el empleo son consecuencias de la serie de inversiones sucesivas.

[6] Para un desarrollo sistemático, mucho más detallado y, además, centrado en el problema del crecimiento, ver Lorente (2018).

También debemos contar con una verdadera competencia dinámica entre empresas, donde la mayor productividad de las nuevas inversiones determina la obsolescencia de una parte de los equipos y puestos de trabajo asociados al capital que fue contabilizado años atrás. De ahí que cada año haya una reducción del capital contable agregado, no porque todas las empresas descarten equipos, sino porque una fracción de ellas se ve obligada a hacerlo. Por esa misma razón, todas las demás empresas necesitan formar reservas de recursos propios para financiar inversiones y conversiones tecnológicas futuras. Estas reservas se forman con una parte de los ingresos de la empresa que permanecen en ella como fondos de depreciación, pero se contabilizan como un costo para que no sean repartidos entre la empresa y sus propietarios, como sucede con el resto del margen sobre costos que permanece como utilidades.

El capital contable aparece entonces como un simple residuo de decisiones de inversión anteriores y sirve para medir la eficiencia de la empresa en comparación con otras, pues los indicadores de rentabilidad convencionales funcionan como una referencia de capacidad de pago y como una medida de eficiencia relativa para los propietarios y para los eventuales financiadores. Pero el papel de este capital contable es estrictamente distributivo: no actúa como factor que genera producto pues, aunque formalmente represente equipos e instalaciones, no existe proporcionalidad con el producto ni es su causa directa.

Se produce lo que se puede vender, y cada inversión decidida con ese propósito es un paquete que asocia costos de equipos (capital) y puestos de trabajo (empleo).

Si queremos medir la capacidad productiva debemos medir cómo cambia la productividad del trabajo con cada nueva inversión, es decir, la relación producto sobre empleo, pues ese es el indicador que guía la decisión de invertir: primero, porque cada proyecto debe ser más rentable que las técnicas en uso para que sea competitivo y, segundo, porque el costo del trabajo es el principal costo de cada empresa y la productividad del trabajo resume sus posibilidades de generar ganancias mejor que cualquier otro índice.

Nuestra teoría contiene entonces relaciones causales que van de (ventas=producto) a una inversión (formación bruta de capital y creación de empleo), que redunda en capital contable (medido en unidades perfectamente homogéneas y aditivas) y en empleo agregado.

En fórmulas, tendríamos una primera cadena cerrada:

$$\Delta Y \rightarrow FBK \rightarrow (-\Delta K \cap + \Delta L) \rightarrow \Delta Y \quad (14)$$

pero aparecen luego otras más, que se entrecruzan en una red compleja:

1. El aumento de productividad de las inversiones nuevas ocasiona la obsolescencia de una parte del capital físico anterior y de los conocimientos de los trabajadores.

2. La necesidad adicional de trabajadores con calificaciones específicas tropieza finalmente con la disponibilidad de personal y, cuando el desempleo es bajo, obliga a las empresas a competir entre sí por los trabajadores, aumentando los salarios y la participación del trabajo en el producto.

3. El aumento de los costos laborales se traduce en mayores ventas en todos los sectores que atienden el consumo masivo y tiende a prolongar los auges, pero cada empresa buscará inversiones más intensivas en equipos y menos en trabajo, frenando los costos, aunque sacrificando parte del aumento potencial de la productividad laboral.

4. El cambio de énfasis en la tecnología preferida por las empresas frena el alza de salarios y de la participación del trabajo, estabilizando la distribución del ingreso, pero frenando el crecimiento de las ventas en todos los sectores de consumo masivo.

5. Y así sucesivamente, formando nuevas cadenas causales que forman una red compleja, pero que admiten cierto grado de homeostasis que se traduce en restricciones, de cumplimiento más estricto cuando son de origen institucional (como la presupuestal que limita el endeudamiento) y algo más laxas cuando dependen de una secuencia de decisiones ajenas que, a veces, se sincronizan parcialmente para generar periodos de auge o de recesión.

A diferencia de la interpretación convencional de factores productivos, que termina explicando las fluctuaciones económicas como consecuencia de choques exógenos de carácter estocástico, esta explicación alterna admite fluctuaciones endógenas, que no son periódicas porque la corriente continua de innovaciones lo impide, pero sí recurrentes. La red compleja de relaciones entre empresas, propietarios y trabajadores se traduce también en relaciones no lineales; y esta clase de sistemas puede tener atractores extraños que combinan cierto grado de determinismo con la imposibilidad de un pronóstico de largo plazo.

CONCLUSIÓN

Comenzamos con una identidad contable y mostramos que se puede convertir en el germen de una explicación teórica al agregar una interpretación de tipo causal. La teoría hoy más difundida escoge

una interpretación que pretende deducir la cantidad de producto a partir de la disponibilidad de factores, especialmente de capital, para descubrir luego que las condiciones de agregación que debe cumplir son demasiado restrictivas.

Las dificultades no nacen de las identidades ni de su conversión en relaciones funcionales, ya que la identidad expresa una relación necesaria que sus variables cumplen indefectiblemente. El problema está en la interpretación causal escogida, que va de la dotación de capital y trabajo a la determinación de la cantidad de producto.

Los problemas de agregación de tipos de trabajo son relativamente simples, porque todo el trabajo ocurre en el mismo periodo: es imposible guardar trabajo y aplicarlo luego porque el tiempo de vida no se traslada ni transfiere. Podemos aceptar entonces alguna convención que "reduzca trabajo compuesto a simple", así sea usando como factor de conversión las diferencias salariales que existan en un momento dado.

No sucede lo mismo con el capital, pues no es posible reducir un tipo de máquina a otro y, además, las máquinas se usan en periodos sucesivos, pero tienen diferencias evidentes con los nuevos diseños, que incorporan mejoras y obligan a descartar las anteriores por simples razones de obsolescencia.

Por otra parte, el capital no aparece en las identidades contables como bienes específicos que requieren un número fijo de trabajadores de determinada calificación y permiten procesar una cantidad específica de insumos para obtener cierta cantidad de producto. Estas son las propiedades que respaldarían la interpretación causal implícita en la teoría neoclásica, pero es claro que la construcción contable del capital solo combina unidades monetarias, aceptando el precio vigente cuando se incorpora el bien de capital a las empresas y aplicando luego un factor de depreciación que, en rigor, mide el impacto de la obsolescencia, ya que el bien específico de capital no se remplaza por desgaste físico sino mucho antes, porque hay mejores equipos disponibles o porque aparecieron productos nuevos.

La restricción entre producción y ventas que presentamos al comienzo de este artículo no es compatible con el supuesto de un capital-factor dotado de una productividad que le es inherente y que se traduce necesariamente en productos, ya que no se produce porque es posible producir y hasta agotar la capacidad física de los equipos disponibles, sino que se produce porque es posible vender.

El capital instalado puede ser la consecuencia de un error de pronóstico de ventas y desaparece de la contabilidad social cuando la

empresa que lo instaló quiebra. Por otra parte, cuando sobreviene una drástica innovación técnica, podría suceder que un capital contable menor sirva para obtener una cantidad de producto mayor, es decir, que, en principio, sería posible producir más con menos capital.

Todo esto sugiere que no existe una relación causal directa de *K* a *Y*, sino una conexión indirecta mediada por otras variables y condicionada por circunstancias contingentes.

Podemos generalizar esta observación a cualquier intento de convertir una identidad en una relación funcional: la relación existe y es exacta, pero no necesariamente como relación causal entre las variables que figuran en ella.

La razón es que las ecuaciones y los modelos formales admiten infinitas interpretaciones, a veces en la misma área de conocimiento y a veces incluso en áreas aparentemente dispares. Un ejemplo trivial es que podemos hallar funciones exponenciales y procesos de crecimiento exponencial en economía, pero también en biología, en química, en astronomía, en física atómica y en muchos contextos más.

Cabe expresar esta misma idea en términos econométricos señalando que para cada conjunto de datos podemos diseñar infinitos modelos diferentes entre sí, cada uno de los cuales expresa una teoría diferente porque el contenido explicativo está en la interpretación de los símbolos formales y no en el formalismo, ni en las correlaciones estadísticas.

Por igual razón, no podemos pensar que sea posible extraer teorías de los datos, aplicando quizá alguna receta inductiva o algún criterio general, como el del modelo más simple posible, porque los mismos datos pueden justificar varias interpretaciones discordes y porque los modelos más simples a veces descartan variables o procesos esenciales que, por simple azar, no son evidentes en el conjunto finito de datos de que disponemos.

Las teorías son hipótesis que arriesgamos para explicar hechos y relacionarlos con otros. Los datos luego podrán confirmarlas o refutarlas.

Y si queremos obtener una teoría que guíe la acción, bien sea para diseñar un experimento o para trazar una política macro, necesitaremos teorías que incluyan hipótesis causales además de las relaciones funcionales, porque la causalidad es lo que convierte la relación en regla de conducta o de política, tan sometida a comprobación como el resto del modelo o teoría.

Las principales objeciones contra la causalidad son que siempre existen varias explicaciones causales para un mismo fenómeno, y que

pueden existir cadenas causales que se cierran sobre sí mismas y se entrelazan en una red compleja. Pero vimos que la proliferación de causas desaparece cuando precisamos el contexto, y vimos también que el enmarañamiento causal se convierte en una ventaja cuando genera alguna forma de estabilidad dinámica que podemos describir como efecto de alguna restricción.

La búsqueda de explicaciones en economía necesita dejar atrás los requisitos doctrinales de óptimos irrealizables y la obsesión por los equilibrios de inspiración estática. Un mundo en cambio permanente solo puede admitir explicaciones basadas en otros cambios previos, en la asimetría de las decisiones y en la interacción entre los individuos que deciden o reaccionan inspirados por intereses comunes de grupo.

Es necesario comenzar por modelos que representen procesos dinámicos, que acepten la posibilidad cotidiana del error y admitan su corrección *a posteriori*, y, sobre todo, que sometan a comprobación sus premisas e hipótesis de partida.

REFERENCIAS BIBLIOGRÁFICAS

Bunge, M. (2009). *Causality and modern science*. New Brunswick: Transaction Publishers.

Candy, J. (1988). *Signal processing: The modern approach*. Nueva York: McGraw-Hill.

Chiarella, C., Flaschel, P. y Franke, R. (2005). *Foundations for a disequilibrium theory of the business cycle: Qualitative analysis and quantitative assesment*. Nueva York: Cambridge University Press.

Corbett, J. y Jenkinson, T. (1997). How is investment financed? A study of Germany, Japan, the United Kingdom and the United States. *Manchester School*, *65*(S-Supplement), 69-93.

Currie, L. y Sandilands, R. (1997). Implications of an endogenous theory of growth in Allyn Young's macroeconomic concept of increasing returns. *History of Political Economy*, *29*(3), 413-443. Trad. al español: Currie, L. y R. Sandilands. (2013). Implicaciones de una teoría del crecimiento endógeno en el concepto macroeconómico de rendimientos crecientes de Allyn Young", *Revista de Economía Institucional*, *15*(28), 95-126.

Duesenberry, J. S. (1949). *Income, saving and the theory of the consumer*. Cambridge: Harvard University Press.

Eberhardt, F., Hoyer, P. y Scheines, R. (2010). Combining experiments to discover linear cyclic models with latent variables. *Journal of Machine Learning*, 9, 185-192.

Felipe, J. y McCombie, J. (2007). Is a theory of total factor productivity really needed? *Metroeconomica*, *58*(1), 195-229.

Fisher, F. (1971). Aggregate production functions and the explanation of wages: A simulation experiment. *Review of Economics and Statistics*, *53*(4), 305-325.

Fisher, F. M. (1969). The existence of aggregate production functions. *Econometrica*, *37*(4), 553-577.
Frank, R. H. y Levine, A. S. (2007). Expenditure cascades. [www.aeaweb.org/annual_mtg_papers/2007/0107_1300_0202.pdf].
Friedman, M. (1953). The methodology of positive economics. En M. Friedman, (ed.), *Essays in positive economics* (pp. 3-43). Chicago: University of Chicago Press.
Gnedenko, B. (1962). *Theory of probability*. Nueva York: Chelsea.
Hackethal, A. y Schmidt, R. (2003). Financing patterns: mesurement, concepts and empirical results. *Finance & Accounting*, Working paper 125.
Harcourt, G. (1972). *Some Cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.
Knuth, D. (1969). *The art of computer programming*. Massachusetts: Addison-Wesley.
Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Lorente, L. (2018). *Dinámica del crecimiento económico*. Bogotá: Universidad Nacional de Colombia.
Lucas, R. (1990). Why doesn't capital flow from rich to poor countries? *American Economic Review*, *80*(2), 92-96.
Marshall, A. (1920). *Principles of economics*. Londres: McMillan and Co.
Mill, J., 1848. *Principles of Political Economy*. London: John Parker.
Nagel, E. (1961). *The structure of science: Problems in the logic of scientific explanation*. Nueva York: Harcourt Brace & World.
Ricardo, D. (1821). *On the principles of political economy and taxation*. Londres: John Murray.
Russell, B. (1917). On the notion of cause. En B. Russell, (ed.), *Mysticism and logic and other essays* (pp. 180-208). Londres: George Allen and Unwin.
Samuelson, P. (1962). Parable and realism in capital theory: The surrogate production function. *Review of Economic Studies*, *29*(3), 193-206.
Solow, R. M. (1956). A contribution to the theory of economic growth. *Quarterly Journal of Economics*, *70*(1), 65-94.
Spirtes, P., Glymour, C. y Scheines, R. (2000). *Causation, prediction, and search*. Cambridge: MIT Press.
Sylos-Labini, P. (1991a). I mutamenti di lungo periodo nei meccanismi che regolano salari e prezzi e il processo di sviluppo. *Rivista di Storia Economica*, 8, 1-26.
Sylos-Labini, P. (1991b). The changing character of the so-called business cycle. *Atlantic Economic Journal*, *19*(3), 1-14.
Veblen, T. (1898). Why is economics not an evolutionary Science?. *Quarterly Journal of Economics*, *12*(4), 373-397.
Veblen, T. (1899). *The theory of the leisure class*. Nueva York: Macmillan.
Von-Wright, G. (1971). *Explanation and understanding*. Londres: Routledge and Kegan Paul.
Young, A. (1928). Increasing returns and economic progress. *The Economic Journal, 38*(152), 527-542. Trad. al español: Young, A. A. (2009). Rendimientos crecientes y progreso económico, *Revista de Economía Institucional*, *11*(21), 227-243.