PARTE III

Anexo estadístico y metodológico

CAPÍTULO 7

Anexo Metodológico

Introducción

Uno de los principales objetivos del *Informe Estado de la Nación* es proveer información oportuna, que permita conocer el avance del país en el logro de sus aspiraciones de desarrollo humano sostenible. En su preparación cada año interviene una amplia red de instituciones e investigadores, quienes colaboran con el suministro de datos actualizados y la aplicación de técnicas y mediciones novedosas, que facilitan una comprensión más objetiva de la realidad nacional. Con la incorporación de distintos instrumentos metodológicos se pretende dar una sólida base técnica a los hallazgos presentados en cada capítulo.

En este Anexo Metodológico se exponen los procedimientos seguidos para el abordaje de algunos temas incluidos en esta edición. Para los capítulos "Equidad e integración social" y "Oportunidades, estabilidad y solvencia económicas" se sintetiza la metodología seguida en la estimación de un modelo de predicción de la pertenencia a un sector económico y la posible movilidad laboral si se cambian algunas características de las personas, así como el impacto que tendría sobre el bienestar de los hogares el ingreso promedio que podrían percibir en el nuevo trabajo.

Otros cinco temas corresponden al capítulo "Oportunidades, estabilidad y solvencia económicas", a saber: i) modelo de los factores determinantes del salario por hora, ii) análisis de los encadenamientos productivos, iii) estudio de la productividad laboral en Costa Rica, iv) perfil sociodemográfico de la moral fiscal, y v) deficiencias en el diseño del impuesto sobre la renta de las empresas.

Para el capítulo "Armonía con la naturaleza" se incluye el procedimiento seguido para valorar la labor que realiza la Secretaría Técnica Nacional Ambiental (Setena) en las evaluaciones de impacto ambiental (EIA) y la efectividad de los instrumentos que se utilizan en ese proceso. También se expone la metodología de construcción de un índice agregado que valora la presencia o no de prácticas sostenibles o amigables con el ambiente en las fincas del país, con datos del VI Censo Nacional Agropecuario, de 2014.

A su vez, el capítulo "Fortalecimiento de la democracia", aporta un análisis de redes conceptuales aplicado a la oferta programática de los partidos políticos y una aproximación novedosa al tema de las acciones colectivas, mediante un estudio de series de tiempo. También se explican los detalles de un análisis de supervivencia realizado con el objetivo de identificar factores que agilizan el trámite legislativo, ya sea reduciendo los tiempos de aprobación o aumentando la probabilidad de que un proyecto se convierta en ley.

Finalmente, para el capítulo "El descontento ciudadano y sus implicaciones para la estabilidad política en Costa Rica", se describen los aspectos técnicos de la encuesta de cultura política "Barómetro de las Américas" del 2015. También se expone el método aplicado en una serie de entrevistas efectuadas con el propósito de entender mejor el fenómeno del descontento ciudadano.

Aportes metodológicos en materia de equidad e integración social

Simulaciones de movilidad laboral

Según la Encuesta Nacional de Hogares (Enaho) del 2015, el mercado laboral costarricense tiene aproximadamente 2.077.348 personas ocupadas. Características sociodemográficas como sexo, nivel educativo, zona de residencia, titulación y dominio de un segundo idioma, entre otras, determinan los sectores económicos en los cuales se inserta esta población y el ingreso que percibe.

Con el objetivo de precisar la relación entre esas características y la pertenencia a sectores económicos específicos, así como la posible movilidad laboral entre ellos, Segura (2016) construyó varios modelos de predicción con las variables de la Enaho 2015. Como punto de partida, se consultaron dos investigaciones, realizadas por Meneses y Anda (2015) y Jiménez-Fontana y Segura (2015). Los modelos aportaron insumos para los capítulos "Equidad e integración social" y "Oportunidades, estabilidad y solvencia económicas".

Las simulaciones se realizaron con técnicas en minería de datos. El método predictivo[1] requiere contar con un conjunto de datos para "entrenar y evaluar" el modelo, es decir, sobre esos datos se prueban y ajustan los parámetros para obtener los resultados más precisos. Para llevar a cabo este proceso se utilizaron las Enaho de 2013 y 2014. Una vez construido el modelo, se efectuaron las predicciones con la información de la Enaho 2015.

El trabajo incluyó la recodificación de algunas variables y la construcción de otras. Por ejemplo, el nivel educativo, el dominio de un segundo idioma y la titulación se combinaron para generar el indicador "capacidades", conformado por cinco categorías: secundaria incompleta sin segundo idioma, secundaria académica completa sin segundo idioma, técnico titulado o segundo idioma sin título académico, tres o más años de universidad sin segundo idioma y tres o más años de universidad con dominio de un segundo idioma.

La edad fue recodificada para generar seis grupos decenales. La región y la zona de residencia se combinaron en una sola variable. Además, dado que la estructura de los hogares influye en la decisión de insertarse y/o permanecer en un sector económico, según condiciones como tradición, patrimonio, clima educativo, entre otros, se crearon seis variables con el porcentaje de importancia de cada sector en el hogar (cantidad de ocupados en el sector *i* con respecto al total de miembros del hogar).

Para determinar el mejor modelo para el conjunto de datos disponible, se aplicaron siete métodos predictivos, denominados: Naive-Bayes, máquinas de soporte vectorial, árboles de decisión, boques aleatorios, métodos de potenciación ADA-Bosting, redes neuronales y el vecino más cercano (kknn). En cada caso se aplicó el siguiente modelo:

Sector económico = sexo+ edad + migrante + central + educnoreg + regzona + aseg + computadora + internet + capacidades + estragro + estrit + estrni + estrserv + estrgob + estrinformal

Donde:

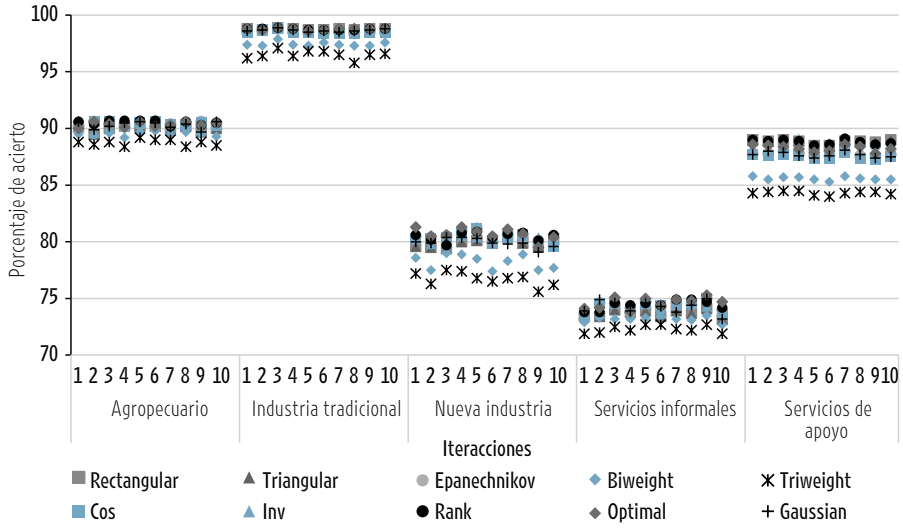
- **sexo** identifica si la persona es mujer u hombre,
- **edad** es una variable recodificada en grupos decenales,
- **migrante** es una variable dicotómica que distingue a las personas según su lugar de nacimiento: Costa Rica u otro país,
- **central** diferencia entre los residentes en la Región Central y los del resto del país,
- **educnoreg** identifica si las personas tienen o no algún tipo de educación no regular,
- **regzona** corresponde a la combinación de región y zona,
- **aseg** identifica a las personas aseguradas y no aseguradas,
- **computadora** e **internet** indican si la persona tiene o no acceso a ambas tecnologías en la vivienda, y
- las variables restantes, **estragro**, **estrit, estrni**, **estrserv**, **estrgob** y **estrinformal** hacen referencia a la estructura interna del agro, industria tradicional, nueva industria, servicios, gobierno y servicios informales, respectivamente.

Cada método fue calibrado y sometido a "validaciones cruzadas", técnica que permite que el modelo se repita "n" veces, omitiendo en cada iteración uno de los "k grupos" de registros previamente indicado. De esta forma es posible visualizar cuan robustas son las estimaciones y, sobre todo, valorar cuan exactas son las estimaciones de la población objetivo. En el mismo proceso se evalúan distintas variaciones de los parámetros y los algoritmos propios del método, a fin de lograr la mejor calibración del instrumento. Para cada método se efectuaron diez iteraciones con cinco grupos para la "validación cruzada". Se utilizaron aproximadamente 31.423 casos para "entrenar" los modelos y alrededor de 15.839 casos para "predecir" el sector económico de las personas ocupadas. Siempre se contó con la variable "real" de los sectores económicos, por lo que las predicciones fueron evaluadas en todo momento. Las rutinas fueron programadas en los paquetes estadísticos R y Stata.

En total se realizaron 250 iteraciones, a partir de las cuales se determinó que el método que mejor predice los sectores económicos para el año 2015 es el "vecino más cercano". El gráfico 7.1 muestra los porcentajes de acierto para cada sector. Las estimaciones para el "gobierno" son los más altos, y de hecho su predicción es casi perfecta (98% en promedio). La "industria tradicional" y la "nueva industria" fueron los más difíciles de predecir; en promedio los aciertos fueron de 80% y 74%, respectivamente.

GRÁFICO 7.1

Estimación del porcentaje de acierto de los sectores económicos, por número de iteraciones, según tipo de algoritmo del método del "vecino más cercano"



Fuente: Segura, 2016.

Como se dijo, el objetivo de la investigación fue aplicar el modelo a poblaciones específicas y determinar sus posibilidades reales de movilización laboral, así como los efectos inmediatos en términos de remuneraciones e impacto sobre la pobreza y la desigualdad de ingresos. Con ese propósito se diseñaron once escenarios, que se describen en el cuadro 7.1. Para cada simulación se elaboraron once bases de datos a las cuales se aplicó el modelo predictivo. Los impactos en las remuneraciones, la pobreza y la desigualdad se evaluaron mediante la imputación del ingreso promedio que podrían percibir los trabajadores en el sector económico predicho. A las personas ocupadas sin secundaria o sin dominio de un segundo idioma se les asignó el promedio de sus contrapartes, es decir, secundaria completa o dominio de un segundo idioma. Una vez imputados los ingresos, se reconstruyó la estructura del ingreso familiar según la metodología del INEC y se recalcularon los principales indicadores de pobreza y distribución del ingreso.

PARA MÁS INFORMACIÓN SOBRE
SIMULACIONES DE MOVILIDAD LABORAL
véase Segura, 2016, en
www.estadonacion.or.cr

Aportes metodológicos en materia de oportunidades, estabilidad y solvencia económicas

Modelo de los factores determinantes del salario por hora

Para identificar las diferencias en el salario por hora entre trabajadores, Meneses y Anda (2015) utilizaron los datos históricos de las encuestas de hogares del INEC desde 1991. El modelo para estimar las diferencias se basa en la ecuación de salarios propuesta por Mincer (1974), que se define como:

$$E\left(\ln(y)|\boldsymbol{x}\right) = \boldsymbol{x}\beta + u$$

En el modelo, la variable y es el salario por hora de los ocupados menores de 65 años y $\boldsymbol{x}$ es un vector de factores determinantes del salario asociados al capital humano, características del empleo y aspectos sociodemográficos. Por su parte, u recoge el efecto promedio de todas las particularidades no observadas que configuran el salario esperado de un trabajador. El cuadro 7.2 presenta los resultados del modelo para el año 2015.

PARA MÁS INFORMACIÓN SOBRE
CRECIMIENTO ECONOMICO Y EMPLEO
véase Meneses y Anda, 2016, en
www.estadonacion.or.cr

Análisis de los encadenamientos productivos

Con el objetivo de examinar los encadenamientos productivos en Costa Rica, Meneses y Anda (2016) analizaron las interconexiones entre industrias y productos a partir de la tipología sectorial de Rasmussen. La fuente de información es la matriz insumo-producto del año base 2012, del Banco Central, la cual contiene datos de 183 productos desagregados según régimen de producción (definitivo o especial).

La tipología de Rasmussen clasifica la producción a partir de dos indicadores. El primero, denominado poder de dispersión (PD), mide la capacidad de un sector para demandar insumos intermedios de otros sectores; es decir, estima la capacidad de "arrastre" de una determinada rama de actividad. Un valor de PD mayor a 1 implica que la rama está altamente relacionada con el sistema económico, pues un crecimiento en su demanda final se expande a todo el aparato productivo más que el promedio y estimula al conjunto de la economía. En contraste, un PD menor a 1 significa que la actividad tiene un poder de dispersión bajo, por lo que un crecimiento en su demanda tiene un débil impacto en el resto de la producción.

El segundo índice se denomina sensibilidad de dispersión (SD) y calcula la

CUADRO 7.1

Escenarios, población objetivo y simulaciones aplicadas

Escenarios	Población objetivo	Simulación
1	Personas desempleadas	Se aplica el modelo al 100% de esta población
2	Personas fuera de la fuerza de trabajo (inactivas)	
3	Personas ocupadas con secundaria incompleta	Se cambia su nivel educativo a secundaria completa con título
4	Personas desocupadas con secundaria incompleta	
5	Personas inactivas de entre 25 a 64 años de edad con secundaria incompleta	
6	Personas ocupadas que no dominan un segundo idioma	Se asume que la persona domina un segundo idioma
7	Personas desocupadas que no dominan un segundo idioma	
8	Personas fuera de la fuerza laboral, de entre 25 y 64 años, que no dominan un segundo idioma	
9	Mujeres jefas de hogar de entre 25 y 64 años de edad, sin pareja, con hijos y fuera de la fuerza laboral	Se aplica el modelo al 100% de esta población
10	Mujeres jefas de hogar de entre 25 y 64 años de edad, sin pareja, con hijos, fuera de la fuerza laboral y con secundaria incompleta	Se cambia su nivel educativo a secundaria completa con título
11	Mujeres jefas de hogar de entre 25 y 64 años de edad, sin pareja, con hijos, fuera de la fuerza laboral y que no dominan un segundo idioma	Se asume que la personas domina un segundo idioma

Fuente: Segura, 2016.

CUADRO 7.2

Modelo de factores determinantes del salario por hora

Variables	2015	Significancia[a/]
Años de escolaridad	0,03	***
Calificado	-1,07	***
Escolaridad postsecundaria	0,10	***
Cursos de educación no regular	0,02	
Segundo idioma	0,19	***
Educación pública	-0,03	*
Edad	0,04	***
Edad al cuadrado	0,00	***
Cuenta propia	-0,21	***
Empleador	0,28	***
Sector público	0,32	***
Sector formal	0,16	***
Sector agropecuario	-0,07	**
Manufactura	-0,20	***
Hombres en sector manufactura	0,23	***
Nueva economía	0,03	
Servicios de apoyo	-0,02	
Empresa grande	0,13	***
Zona urbana	0,07	***
Unido	0,10	***
Hombre	0,08	***
Constante	5,84	***
r^2	0,44	
N	13.940	

a/ Niveles de significancia: * p<0.05; ** p<0.01; *** p<0.001.

Fuente: Meneses y Anda, 2016.

CUADRO 7.3

Tipología sectorial según Rasmussen

		Poder de dispersión	
		< 1	>= 1
Sensibilidad de dispersión	> = 1	Estratégico	Clave
	< 1	Independiente	Impulsor

Fuente: Meneses y Anda, 2016, con base en Schuschny, 2005.

posibilidad de que los bienes finales de un sector sean utilizados como insumo por otras actividades; es decir, cuantifica la capacidad de un sector para "empujar" la economía. Si la SD es mayor a 1, el estímulo generado por el incremento en la demanda final del conjunto de las actividades productivas es superior al promedio. Por otro lado, un índice de SD menor a 1 indica que la actividad es menos sensible a cambios generales en la demanda. Los resultados de estos índices se utilizan para clasificar los sectores económicos en los cuatro grandes grupos que muestra el cuadro 7.3.

Estudio de la productividad laboral en Costa Rica

Mulder et al., 2016 estudiaron la productividad laboral en Costa Rica, mediante un proceso que consta de tres partes. Primero, la productividad se estima como la relación entre el PIB y la población ocupada para los países centroamericanos, y este indicador es normalizado con respecto al nivel de Estados Unidos. Para realizar este cálculo se utilizó la base de datos de indicadores de la OIT.

En una segunda etapa se estima la productividad laboral como la relación entre el valor agregado y la población ocupada por actividad económica, a partir de la matriz insumo-producto (MIP) del año base 2012 (BCCR). Las ramas de la MIP fueron agregadas a 35 sectores según la clasificación internacional CIIU 4. El análisis solo consideró los sectores con una participación superior al 0,1% en el empleo total. Debido a limitaciones en los datos, la variable de productividad utilizada fue la productividad laboral y no tomó en cuenta otras variables, como el capital.

Finalmente, en la tercera fase se realiza un análisis de descomposición, con el objetivo de determinar los factores que impulsan o limitan el crecimiento de la productividad en el tiempo. La descomposición se desagregó para diez sectores en los períodos 2001-2008 y 2008-2015. Las fuentes de información fueron los registros de cuentas nacionales con año base 1991 para las cifras de producción, y los datos de desempleo se obtuvieron de las encuestas de hogares del INEC. La metodología de la descomposición se basa en las ecuaciones de TeeWei (2013), las cuales permiten determinar si los cambios obedecen a un mejor desempeño a lo interno de los sectores, o más bien a una reasignación de los trabajadores a sectores con distinta productividad laboral (cuadro 7.4). Este ejercicio permite identificar tres tipos de efectos, a saber:

- **Efecto interno:** indica el aumento de la productividad al interior de los diversos sectores, es decir, la contribución del crecimiento de cada actividad al cambio de la productividad total.

- **Efecto estructura o efecto intersectorial estático:** muestra los cambios en la participación sectorial del empleo, esto es, el aporte de las variaciones en la proporción del empleo de los sectores con distintos niveles de productividad al crecimiento general de la productividad.

CUADRO 7.4

Componentes de la ecuación de descomposición de la productividad

Ecuación/variable	Detalle
$\frac{P_t - P_{t-1}}{P_{t-1}} = C_1 + C_2 + C_3$	Indica la variación relativa de la productividad.
$C_1 = \sum_{i=1}^{n} \left[\left(\frac{P_t - P_{t-1}}{P_{t-1}} \right) x \frac{Y_{it-1}}{Y_{t-1}} \right]$	Efecto cambio interno.
$C_2 = \sum_{i=1}^{n} \left[\left(\frac{P_{it-1}}{P_{t-1}} \right) x \left(\frac{L_{it}}{L_t} - \frac{L_{it-1}}{L_{t-1}} \right) \right]$	Efecto cambio estático.
$C_3 = \sum_{i=1}^{n} \left[\left(\frac{P_{it} - P_{it-1}}{P_{t-1}} \right) x \left(\frac{L_{it}}{L_t} - \frac{L_{it-1}}{L_{t-1}} \right) \right]$	Efecto cambio dinámico.
P_t	Es el nivel de productividad de la economía en el período *t* medido como la relación entre el PIB a precios constantes de 1991 y la población ocupada.
$Y = \sum_{i=1}^{n} Y_{it}$	*Y* es el total del PIB de la economía en el período *t* y Y_{it} es el PIB del sector *i* en el año *t*.
$L = \sum_{i=1}^{n} L_{it}$	*L* es el total de la población ocupada (población en edad de trabajar) de la economía en el período *t*, y L_{it} es la población ocupada del sector *i* en el año *t*.
$i = 1, \ldots, n$ $\frac{Y_{it-1}}{Y_{t-1}}$	Se refiere al enésimo sector en la economía. Señala la proporción del PIB sectorial con respecto al PIB total.
$\frac{P_{it-1}}{P_{t-1}}$	Muestra la proporción de la productividad.
$\frac{L_{it}}{L_t} - \frac{L_{it-1}}{L_{t-1}}$	Hace referencia a la variación de la proporción de empleo sectorial.

Fuente: Mulder et al., 2016.

- **Efecto intersectorial dinámico:** identifica la contribución de los cambios en la proporción del empleo de los sectores con diferentes tasas de crecimiento de la productividad, a las variaciones de la productividad total.

PARA MÁS INFORMACIÓN SOBRE **CRECIMIENTO DE LA PRODUCTIVIDAD EN COSTA RICA** véase Mulder et al., 2016, en www.estadonacion.or.cr

La moral fiscal costarricense

Para indagar acerca de la moral fiscal de los costarricenses, Botey (2016) utilizó los datos de la encuesta regional de percepción pública *Latinobarómetro,* en su edición de 2015. Específicamente, empleó como variables independientes dos indicadores: politización ciudadana e insatisfacción con los servicios públicos.

Ambas variables se crean como la suma de una serie de otras variables que se detallan en el cuadro 7.5. Para confirmar que, en efecto, los indicadores representan una sola dimensión, se estimó el alpha de Cronbach, que resultó ser mayor a 0,7 en ambos casos, lo cual indica que los indicadores son coherentes y fiables para medir cada concepto.

Con el fin de determinar el perfil sociodemográfico de la moral fiscal en Costa Rica, se construyeron dos modelos de regresión binomial probit. Las variables dependientes se aproximaron a partir de dos preguntas: i) ¿cuán justificable cree usted que es evadir impuestos? y ii) ¿cuán dispuesto está a que se aumenten los impuestos y/o que el país se endeude para financiar obras de infraestructura que favorezcan la integración con el mundo (puentes, autopistas, aeropuertos, puertos)? Ambas preguntas fueron aplicadas a mil costarricenses y obtuvieron tasas de respuesta de 94,6% y 87,5%, respectivamente.

Para tener un panorama más amplio de los datos, se crearon variables binarias a partir de la escala de valores de las dos preguntas antes presentadas. Se asignó un valor de 0 a los ciudadanos que para nada justifican la evasión y 1 a los que la justifican en alguna medida. Además, se asignó un valor de 0 a quienes aceptan algún tipo o todo el aumento de impuestos necesario para financiar infraestructura y 1 si no aceptan ningún aumento de impuestos para ese fin. El cuadro 7.6 presenta los resultados de los modelos.

CUADRO 7.5

Detalle de indicadores de politización ciudadana e insatisfacción con los servicios públicos

Indicador	Variables	Alpha de Cronbach
Politización ciudadana	Participación en organizaciones políticas, participación en elecciones, participación en protestas sobre temas de educación o salud, participación en marchas autorizadas, participación en marchas prohibidas, realiza reclamo en redes sociales y asiste a reuniones para participar en una petición.	0,711
Insatisfacción con los servicios públicos	Insatisfacción con funcionamiento de policía, hospitales públicos, burocracia, sistema judicial, servicios de electricidad, educación pública y transporte.	0,804

Fuente: Botey, 2016.

CUADRO 7.6

Modelo de moral fiscal

Variables	Modelo 1 Justifica totalmente la evasión		Modelo 2 Total desacuerdo con impuestos para infraestructura	
Variables sociodemográficas				
Sexo (Hombre=0/Mujer=1)	-0,126		0,309	***
Edad	-0,016	***	0,009	***
Pequeña aglomeración	0,030		-0,042	
Gran aglomeración	0,144		0,008	
Educación básica	-0,069		0,325	**
Educación secundaria	0,072		0,194	
Variables políticas y económicas				
Participación en organizaciones sociales	-0,258	***	-0,143	
Politización del ciudadano	-0,032	***	-0,012	
Importancia asignada al medio ambiente			0,192	**
Dificultades económicas del país			0,077	
Variables de calidad del sistema democrático				
Falta de transparencia del gobierno	-0,115	**	0,134	**
La corrupción ha empeorado	-0,107	**	0,089	*
Variables sobre la calidad del gasto y la capacidad redistributiva del sistema				
Insatisfacción con servicios públicos	0,025	**	-0,004	
Distribución injusta	-0,217	***	0,282	***
Constante	1,894	***	-2,592	***
Observaciones	824		822	

Fuente: Botey, 2016.

PARA MÁS INFORMACIÓN SOBRE
CULTURA TRIBUTARIA EN COSTA RICA
véase Botey, 2016, en
www.estadonacion.or.cr

Deficiencias en el diseño del impuesto sobre la renta de las empresas

Con el propósito de analizar las deficiencias en el diseño del impuesto sobre la renta de las empresas, Bachas y Soto (2016) realizaron una estimación microeconométrica de la elasticidad o sensibilidad de las utilidades a la tasa impositiva. Para ello utilizaron el universo de negocios costarricenses que declararon impuestos ante el Ministerio de Hacienda en el período 2008-2014. Esto significó trabajar con datos de 222.352 empresas y un total de 617.929 observaciones a lo largo del los siete años del período. Las empresas estudiadas tienen ventas por debajo de 150 millones de colones y representan el 85% de las 80.000 unidades productivas que declaran impuestos cada año, el 25% de la renta neta total y el 15% de lo recaudado con este tributo. Con base en esa información se estimó la elasticidad de las utilidades a la tasa impositiva, en seis pasos:

1. Se agrupan las observaciones en tramos de 0,5 millones de renta bruta (ventas) con su valor promedio de renta neta (utilidades).

2. Se observa la distribución de estos puntos y se aprecia que hay un exceso de densidad antes de cada umbral. Posteriormente se estima ese exceso.

3. Mediante un método de punto de convergencia (Kleven y Waseem, 2013), se aplica este exceso de densidad a la izquierda del umbral con un faltante de densidad de igual magnitud a la derecha del umbral, que arroja un nivel máximo de renta bruta afectado por el salto de tasa impositiva en el umbral.

4. Este nivel máximo se toma como la variación total de la renta bruta y, junto al cambio de la tasa impositiva (Saez, 2011), se utiliza para estimar la elasticidad de la renta bruta.

5. Se evidencia que después de cada umbral hay una caída en la renta neta por vía de un mayor costo promedio. Este mayor costo se toma como el cambio total del costo a la tasa impositiva y se utiliza para estimar la elasticidad del gasto.

6. Se procede a utilizar la elasticidad de las ventas y el costo para estimar, finalmente, la elasticidad de la renta neta al cambio de la tasa impositiva.

PARA MÁS INFORMACIÓN SOBRE
CUMPLIMIENTO, PROGRESIVIDAD Y RECAUDACIÓN DEL IMPUESTO SOBRE LA RENTA
véase Bachas y Soto, 2016, en
www.estadonacion.or.cr

Aportes metodológicos en materia de armonía con la naturaleza

Sistematización y análisis de los expedientes de la Setena

En reiteradas ocasiones este Informe ha señalado que la evaluación y control del impacto ambiental de las actividades humanas y productivas, por su importancia para la gestión en este campo, constituyen un reto para la sostenibilidad del desarrollo nacional. En esta edición se efectuó un nuevo análisis sobre este tema, con el propósito de valorar la labor que realiza la Secretaría Técnica Nacional Ambiental (Setena) en relación con las evaluaciones de impacto ambiental (EIA) y la efectividad de los instrumentos que se utilizan en ese proceso.

Se aplicó una metodología exploratoria basada en una mezcla de técnicas de investigación cualitativas y cuantitativas. En primera instancia, se creó una base de datos con todos los expedientes ingresados a la Setena en 2014, para un total de 2.288 casos, distribuidos según la actividad económica en que fueron clasificados por esa institución[2], así como por su localización geográfica (provincia, cantón y distrito), tipo de expediente, estado actual del trámite (viabilidad aprobada, rechazada, bajo análisis, archivada, suspendida, con prórroga, con nulidad) y las acciones de seguimiento institucional de las que han sido objeto. Esto generó algunos de los hallazgos que se presentan en el capítulo 4.

La base de datos contiene la información aportada por los desarrolladores de los proyectos, que incluye estudios de factibilidad, el formulario de evaluación ambiental preliminar (FEAP)[3] y las medidas previstas para mitigar los impactos potenciales, así como los documentos que describen la interacción que tiene lugar entre la Setena, otras entidades públicas y los responsables de los proyectos, durante la valoración de las solicitudes.

En cuanto a los tipos de expedientes, existen tres categorías

- Formulario D1: documento de evaluación de ingreso inicial para proyectos de alto y moderado impacto ambiental (niveles A, B1 y B2).

- Formulario D2: documento de evaluación de ingreso inicial para proyectos de bajo impacto ambiental (nivel C).

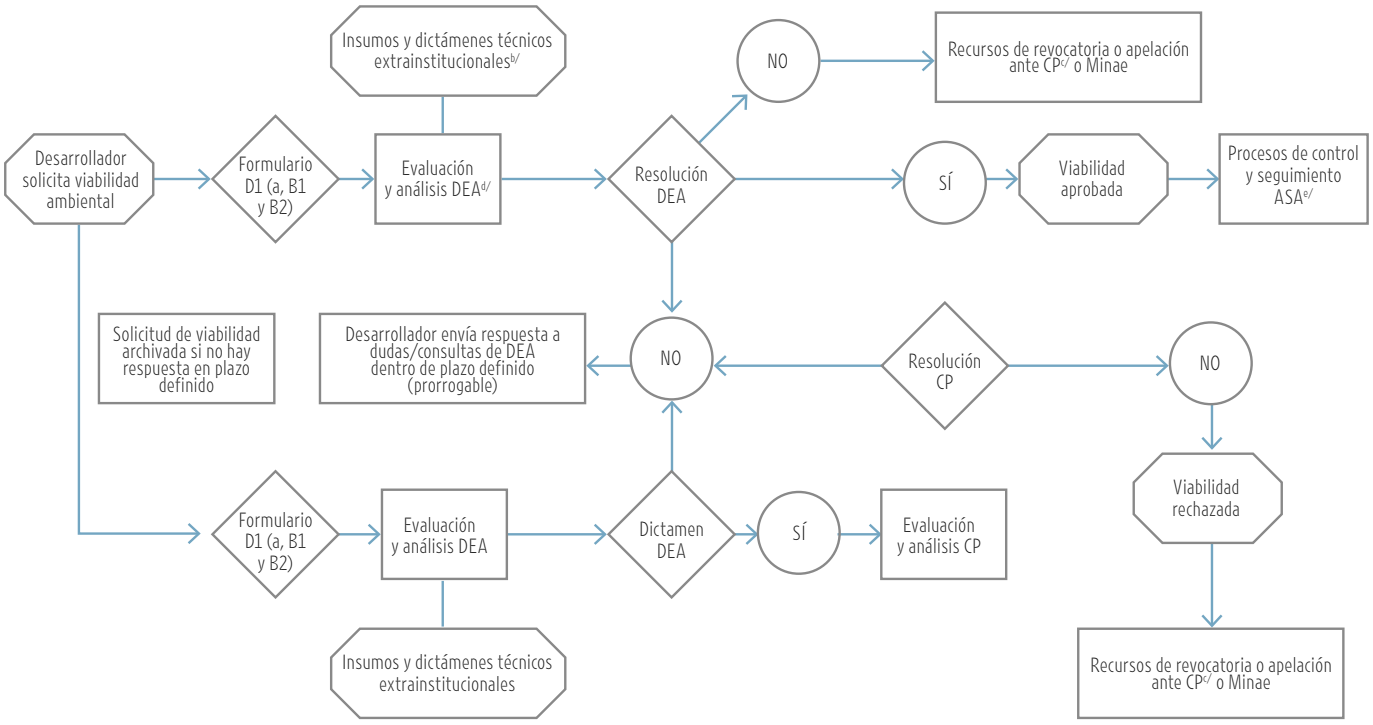
- Estudios de diagnóstico ambiental EDA: instrumento voluntario de evaluación ambiental, aplicable a los proyectos que se desarrollaron sin que, en su momento, la legislación les exigiera algún trámite ante la Setena[4].

Del total de expedientes sistematizados, un 47,5% correspondió a formularios D1 y, de estos, solo un 36,9% había sido aprobado al cierre de edición de este Informe.

Posteriormente se seleccionó una muestra de noventa proyectos de moderado y alto impacto ambiental, tomados de manera aleatoria entre los que ingresaron en 2014 y ya se encuentran aprobados. Con ellos se realizó una evaluación cualitativa centrada en cinco aspectos (diagrama 7.1), a saber: i) la documentación entregada por el desarrollador en su solicitud, ii) los oficios en que la Setena respondió o pidió más detalles, iii) las respuestas de los desarrolladores a esos oficios y los

DIAGRAMA 7.1

Trámite de solicitudes de viabilidad ambiental en la Setena[a/]



a/ Simbología del diagrama: rombo: documento relevante; rectángulo: proceso de valoración o seguimiento; circulo: decisión; octógono: insumos para la decisión.
b/ Incluye, entre otras, las siguientes entidades: Senara, Dirección de Aguas del Minae, Dirección de Geología y Minas, Sinac, Ministerio de Salud, MOPT y municipalidades.
c/ Comisión Plenaria.
d/ Departamento de Evaluación Ambiental.
e/ Departamento de Auditoría y Seguimiento Ambiental de Setena.

estudios aportados (si los hubo), iv) las resoluciones finales del Departamento de Evaluación Ambiental y la Comisión Plenaria de la Secretaría y v) los reportes elaborados por el Departamento de Auditoría y Seguimiento Ambiental de la institución.

Por último, se contrastó la información del FEAP y los estudios técnicos con la documentación producida por la Setena, a fin de determinar cuáles aspectos de la valoración de impactos realizada por el desarrollador son validados por la institución.

PARA MÁS
DETALLES METODOLÓGICOS
véase Ramírez, 2016, en
www.estadonacion.or.cr

Índice de prácticas sostenibles en las fincas agropecuarias

En el marco del análisis sobre el uso agropecuario del territorio, se construyó un índice agregado que valora la presencia o no de prácticas sostenibles o amigables con el ambiente en las fincas del país. Para ello se utilizó la información recabada por el VI Censo Nacional Agropecuario, que llevó a cabo el Instituto Nacional de Estadística y Censos (INEC) en 2014.

En primera instancia se identificaron y agruparon por áreas temáticas las 35 variables relacionadas con las prácticas productivas sostenibles que se detallan en el cuadro 7.7, y que corresponden a los esfuerzos desplegados por las fincas entre mayo de 2013 y abril de 2014. Posteriormente, con el criterio de expertos se ordenaron las variables según su importancia en términos de sostenibilidad y se les asignó un puntaje.

Dado que el Censo registra cada una de las prácticas por producto, fue necesario obtener puntajes en ese nivel de desagregación y, para obtener un valor resumen de estos últimos, se trabajó con un promedio simple. Al final de cada proceso se obtuvieron cuatro constructos, uno para cada área temática, los cuales fueron re-escalados para que sus valores se movieran entre 0 (ausencia de prácticas sostenibles) y 10 (alta presencia de prácticas sostenibles), lo que permitió realizar ejercicios de interpretación y comparación.

Un segundo paso consistió en obtener un puntaje global, que resumiera el índice por fincas. En este caso también se aplicó un promedio simple de los cuatro constructos, para obtener un valor único ubicado en el rango de 0 a 10 puntos. Por último, para efectos de análisis y procesamiento, se definieron cinco quintiles del

CUADRO 7.7

Agrupación de variables sobre prácticas agropecuarias según área temática. 2014

Área temática	Variables
Cultivos	Sistemas de riego, fertilizantes, control de malezas, fungicidas y otros plaguicidas
Terreno agrícola	Sistema agroforestal, cultivos intercalados, rotación de cultivos, cercas vivas y barreras rompevientos, siembras de contorno, quema controlada y tratamiento de residuos
Uso pecuario	Cercas vivas, certificado veterinario, cercas eléctricas, inseminación artificial, tratamiento de desechos vacunos, tratamiento de desechos de cabras, tratamiento de desechos de ovejas, tratamiento de desechos de cerdos
Otras prácticas	Utilización de drenajes, tratamiento de aguas, energía con páneles solares, energía con biogás, energía con residuos agrícolas, energía con electricidad, energía con generador, energía con otras fuentes, energía con leña, energía con combustible y pago de servicios ambientales

CUADRO 7.8

Distribución de las fincas por de tamaño, según quintil de sostenibilidad

Tamaño de las fincas (hectáreas)	Primer quintil[a/]	Segundo quintil[b/]	Tercer quintil[c/]	Cuarto quintil[d/]	Quinto quintil[e/]	Total
0,00 - 0,49	2.007	1.774	1.934	1.922	1.665	9.302
0,49 - 1,00	2.516	2.059	2.220	2.232	1.964	10.991
1,00 - 1,90	1.627	1.414	1.492	1.627	1.476	7.636
1,91 - 3,00	2.675	2.285	2.227	2.376	2.168	11.731
3,01 - 4,22	1.372	1.278	1.378	1.443	1.378	6.849
4,23 - 6,99	1.897	1.897	1.929	1.894	1.914	9.531
6,99 - 10,48	1.854	1.969	1.778	1.699	1.855	9.155
10,48 - 20,00	1.797	1.994	1.886	1.795	2.010	9.482
20,08 - 50,00	1.608	2.180	1.841	1.896	2.057	9.582
50,11 - 13.021	1.267	1.784	1.872	1.719	2.116	8.758
Total	**18.620**	**18.634**	**18.557**	**18.603**	**18.603**	**93.017**

a/ Fincas que obtuvieron una calificación promedio de entre 0 y 2,52 puntos.
b/ Fincas que obtuvieron una calificación de entre 2,53 y 3,38 puntos.
c/ Fincas que obtuvieron una calificación promedio de entre 3,39 y 4,18 puntos.
d/ Fincas que obtuvieron una calificación de entre 4,19 y 5,06 puntos.
e/ Fincas que obtuvieron una calificación de entre 5,07 y 5,83 puntos.

Fuente: Elaboración propia con datos del INEC, 2015.

índice, y diez deciles de tamaño de finca, tal como se muestra en el cuadro 7.8.

PARA MÁS
DETALLES METODOLÓGICOS
véase González et al., 2016, en
www.estadonacion.or.cr

Aportes metodológicos en materia de fortalecimiento de la democracia

Análisis de redes conceptuales aplicado a la oferta programática de los partidos políticos

Con el objetivo de profundizar en el seguimiento de la oferta programática de los partidos políticos, se emplearon dos técnicas computacionales denominadas "recuperación de Información" (RI) y "análisis de redes sociales" (ARS) para realizar el estudio sobre redes conceptuales. Estas técnicas se aplicaron a los programas de gobierno presentados para las elecciones nacionales de 2014 por los siguientes partidos: Accesibilidad sin Exclusión, Acción Ciudadana, Frente Amplio, Liberación Nacional, Movimiento Libertario, Renovación Costarricense y Unidad Social Cristiana. Además se analizaron las transcripciones de las entrevistas a los candidatos presidenciales de esos partidos, realizadas por el PEN en el transcurso de la campaña electoral.

La técnica de RI consiste en la implementación de algoritmos automatizados

para extraer contenidos relevantes de una colección de documentos. En este caso, se decidió unificar el programa de gobierno y la transcripción de la entrevista en un solo documento para cada partido. Seguidamente se realizó una depuración léxica que implicó la eliminación de signos de puntuación y exclamación, acentos ortográficos, numerales y caracteres especiales, así como la conversión del texto a letras en minúscula. Además se suprimieron las palabras que no poseen carga semántica, como los pronombres y las preposiciones, y otros términos que se consideraron no relevantes, como los verbos "ser" y "estar", entre otros.

Luego se hizo una derivación de los términos, de manera que aquellos semánticamente similares quedaran agrupados por su raíz. Por ejemplo, la raíz "educ" representaría los términos "educación", "educativa", "educadoras". Para esto, se empleó el algoritmo de Porter (1980), según el cual las variaciones morfológicas de la mayoría de las palabras se ubican en los sufijos. Este algoritmo es muy utilizado, debido a que emplea una medida que considera la cantidad de consonantes, vocales y sílabas resultantes después de remover el sufijo; con ello se evitan dos posibles limitaciones del proceso de derivación: la *sobretruncación,* que se da cuando una raíz se aplica a varios conceptos, y la *subtruncación*, que ocurre cuando un término queda fuera de la raíz (Willet, 2006). Por último se obtiene un índice de la frecuencia de cada una de las raíces.

Una vez que se contó con los términos derivados se aplicó el ARS. El propósito de este ejercicio es brindar mediciones que faciliten la descripción y comprensión del comportamiento de las redes sociales. Se parte de la premisa de que estas poseen una estructura que se hace evidente en los patrones regulares de interacción entre las entidades que participan en ellas, por ejemplo personas, grupos pequeños, organizaciones, entre otros (Knoke y Yang, 2008). La red de conceptos se construye mediante algoritmos que sistemáticamente generan relaciones entre las palabras presentes en el texto.

Para aplicar esta técnica se siguió el enfoque de Paranyushkin (2011), en el cual la relación entre los términos se determina según su proximidad en el texto. Así, se recorren todos los términos y a cada uno se le asigna una relación con al menos dos términos próximos. La proximidad, es decir, la cantidad de palabras que se debe recorrer antes de asignar una relación, estará determinada por el tamaño del salto o distancia deseada entre las palabras, según criterio experto. En este caso particular se utilizó una distancia de tres términos, que mostró una mayor capacidad de interpretación del objeto de interés durante las primeras pruebas. Es decir, para cada término de la lista se estableció una relación con los siguientes dos términos que se encontraban a tres palabras de distancia. Si una relación entre términos se repite, se le aumenta el peso, como un indicador de importancia en el discurso de la relación.

Además se utilizó un filtro basado en un conjunto de descriptores cuya ocurrencia en el texto se conoce por sistematizaciones previas de la oferta programática realizadas por el PEN. Estos son: solvencia y eficiencia del Estado; productividad y empleo; desigualdad, pobreza y seguridad social; ambiente y energía, y gestión política. A cada uno se le asignó una serie de palabras clave. De esta forma, la red conceptual estuvo conformada por los términos asociados a cada descriptor aplicando el algoritmo de construcción de relaciones a dos niveles de profundidad. Dicho de otra forma, para cada descriptor se crearon relaciones con sus correspondientes dos términos próximos. Esto permitió generar una lista de nodos y otra de aristas o relaciones. Para determinar la relevancia de cada nodo de la red se aplicó el algoritmo PageRank (Page et. al, 1999), el cual hace un balance entre la cantidad de relaciones que posee el nodo y el peso de cada una, obteniendo así un indicador de importancia relativa del nodo en la red.

Finalmente, para efectos de la visualización de las redes se consideraron diversos parámetros (cuadro 7.9). Se utilizó el programa Gephi y el algoritmo Fruchterman & Reinhold (1991), el cual organiza los nodos en función de su relevancia y toma en cuenta la cantidad de enlaces con los que el nodo se relaciona y el peso asociado a ellos.

CUADRO 7.9

Parámetros para la visualización de redes en la oferta programática de los partidos políticos en Costa Rica

Parámetro	Análisis de la oferta programática
Métrica utilizada para filtrado	Page Rank
Filtro de visualización	2,5% de los nodos
Tamaño de nodos[a/]	Escala 10 - 120 en el factor del Page Rank.
Aristas visibles[b/]	Se efectúan tres representaciones gráficas: i) Aristas visibles, cuando su peso es superior o igual a 3, es decir, cuando la relación entre el par de nodos se refuerza con 3 o más reincidencias en el texto. ii) Aristas visibles, cuando su peso se encuentra en el rango de [x/2 - x], siendo x el peso máximo registrado entre un par de nodos. Es decir, si la relación más fuerte entre un par de nodos tiene 40 reincidencias, en la representación gráfica se incluirán las aristas cuyo peso se encuentra en el rango [20-40]. iii) Sin peso, todas las aristas se grafican como si su peso fuera cero. Como resultado las aristas aparecen ocultas o no visibles.

a/ Se usa el valor de 10 en el factor de Page Rank para graficar los nodos más pequeños y se va aumentando gradualmente hasta llegar al valor de 120 para los más grandes.
b/ En el caso del partido Renovación Costarricense se omitieron las aristas con peso superior a 3, debido que el peso máximo registrado fue 3.

Fuente: Céspedes y Segura, 2016.

PARA MÁS INFORMACIÓN SOBRE
ANÁLISIS DE REDES CONCEPTUALES EN LA OFERTA PROGRAMÁTICA DE LOS PARTIDOS POLÍTICOS
Véase Céspedes y Segura (LIIT-UNED), 2016, en
www.estadonacion.or.cr

Análisis de supervivencia de los proyectos de ley

En esta edición se llevó a cabo un análisis de supervivencia de los proyectos de ley, con el objetivo de identificar los factores que determinan los tiempos de aprobación o la probabilidad de que una iniciativa se convierta en ley. Para ello se construyó una base de datos con detalles de los 6.015 proyectos presentados entre el 1 de mayo de 2000 y el 30 de abril de 2016. Este es un método estadístico enfocado en estudiar los tiempos de ocurrencia de un fenómeno y su relación con variables de interés.

En la aplicación de la técnica se debe contar con una fecha de inicio y una final. En este caso, el momento inicial corresponde al día en que a un proyecto de ley se le asignan un número de expediente y una comisión para su estudio. Se da por concluida una iniciativa por dos motivos: aprobación o archivo; el día en que ocurre uno u otro resultado se considera la fecha final. En el cuadro 7.10 se detallan los dos tipos de variables: categóricas o continuas. En el primer caso la categoría de referencia es aquella que se codifica con el número cero, elegida mediante criterio experto.

El análisis realizado se basó en tres enfoques: el semiparamétrico, el paramétrico y el de fracción de cura (Klein

CUADRO 7.10

Detalle de las variables incluidas en el análisis de supervivencia de los proyectos de ley

Variable	Descripción	Rango
Variables categóricas[a/]		
Iniciativa	Indica si el proyecto de ley fue presentado por el Poder Ejecutivo o el Legislativo.	0: Legislativa 1: Ejecutiva"
Legislatura	Cada período legislativo está compuesto por cuatro legislaturas que van desde el 1 de mayo de un año hasta el 30 de abril del año siguiente.	0: Cuarta 1: Primera 2: Segunda 3: Tercera
Partido presidente en el Congreso	Partido político al que pertenece el Presidente de la Asamblea Legislativa, electo en la legislatura en que fue presentado el proyecto de ley.	0: PUSC 1: PASE 2: PLN 3: PAC
Proponentes[b/]	Partido que propone el proyecto de ley.	0: PLN 1: PUSC 2: PAC 3: Varios 4: Otros
Sesión legislativa	Existen dos tipos de sesiones: las ordinarias, en las que los diputados controlan la agenda legislativa, y las extraordinarias, en las que el Ejecutivo es quien convoca. Cada legislatura está dividida en cuatro períodos: dos de sesiones ordinarias (del 1 de mayo al 31 de julio y del 1 de septiembre al 30 de noviembre) y dos de sesiones extraordinarias (del 1 al 31 de agosto y del 1 de diciembre al 30 de abril del año siguiente).	0: II Ordinaria 1: I Ordinaria 2: I Extraordinaria 3: II Extraordinaria
Consulta constitucional	Indica si se realizó alguna consulta constitucional sobre el proyecto.	0: No 1: Sí
Dispensa de trámite	Aplicación del artículo 177 del Reglamento de la Asamblea Legislativa, el cual indica que un proyecto puede ser conocido en primer debate, sin el requisito del informe previo de una comisión.	0: No 1: Sí
Informes	Indica la existencia de informes sobre el proyecto presentados por cualquier órgano interno de la Asamblea u otras instituciones externas consultadas al respecto.	0: No 1: Sí
Variables continuas		
Dictámenes de fondo	Cantidad total de dictámenes emitidos para cada proyecto por las comisiones permanentes y las comisiones especiales. No se incluyen los dictámenes de la comisión de redacción.	0 a 7
Informes de mociones	Cantidad de informes de mociones presentados al proyecto de ley.	0 a 12

a/ La categoría de referencia es aquella en la que se presentó la mayor cantidad de proyectos y se codifica con el número 0, a excepción de la variable "legislatura", para la cual la referencia fue elegida mediante criterio experto. En el caso de las variables dicotómicas (consulta constitucional, dispensa de trámite e informes), la categoría de referencia es la ausencia del rasgo.
b/ La categoría "varios" refiere a las iniciativas que presentaron en conjunto varios partidos, mientras que la categoría "otros" son los proyectos presentados por un solo partido distinto al PUSC, el PAC o el PLN.

y Moeschberger, 2003). La técnica semiparamétrica utilizada es un modelo de regresión de Cox que permite incluir y evaluar el efecto de variables predictoras sobre la ocurrencia de un evento. En este caso el modelo se usó para estimar el efecto (aumento o reducción) de cada una de las variables consideradas sobre las tasas de aprobación de los proyectos de ley.

Por su parte, el enfoque paramétrico asume una distribución de probabilidad específica; las más utilizadas son Weibull, Exponencial, Gompertz, Log-logística, Log-normal y Gamma. En este análisis se trabajó con una distribución log-logística, para estimar si las variables aumentan o disminuyen el tiempo para la aprobación de un proyecto.

Finalmente, la fracción de cura es un tipo especial de análisis de supervivencia que asume que hay una proporción de individuos que nunca experimentará el evento de interés, lo que en este caso se refiere a los proyectos de ley que no tienen posibilidad alguna de ser aprobados, pues ya fueron archivados.

PARA MÁS INFORMACIÓN SOBRE
ANÁLISIS DE SUPERVIVENCIA DE PROYECTOS DE LEY
Véase Solórzano, 2016, en www.estadonacion.or.cr

Análisis de series de tiempo sobre las acciones colectivas

El PEN tiene una base de datos que diariamente registra las acciones colectivas que ocurren en el país. Esta iniciativa surgió en 2001 como un proyecto de investigación conjunta con el Instituto de Investigaciones Sociales de la UCR. En 2010 el Programa hizo una revisión metodológica y decidió darle a continuidad al trabajo de recolección y registro, con dos objetivos: la actualización anual de la base de datos y la ampliación de la serie de tiempo hacia atrás. Actualmente se cuenta con información para el período enero de 1992-marzo de 2016.

La fuente de información de la base de datos son tres medios de prensa escrita de circulación nacional: *La Nación*, *Diario Extra* y *Semanario Universidad*. Una decisión metodológica fue utilizar las versiones impresas de cada medio para asegurar que la información registrada no sufra modificaciones con el tiempo. Se recopilan las principales noticias sobre acciones colectivas y otras de tipo complementario que enriquecen el estudio. Estas son sistematizadas en una bitácora y posteriormente incorporadas a la base de datos mediante el paquete estadístico SPSS. Para el registro de cada variable se sigue una serie de pasos y definiciones que están consignados en un manual metodológico. Tanto la base de datos como su manual se encuentran a disposición del público en el sitio www.estadonacion.or.cr.

Este año se aprovechó el acervo de información recopilada para realizar un análisis de series de tiempo. La base de datos cumple con los tres requisitos básicos para la aplicación de esta metodología. Primero, es una colección de observaciones (datos) sobre un fenómeno, registradas manera sistemática en el tiempo; en este caso son registros diarios de acciones colectivas, agregados de forma mensual. Segundo, es un set de datos estocásticos, es decir, que evolucionan en el tiempo, de modo que es posible predecir con ciertos grados de precisión su comportamiento futuro, a partir de la información conocida hasta el momento. Tercero, la colección de observaciones supera el período mínimo de cinco años; dado que se cuenta con registros para veinticuatro años, incluso fue posible incluso dividir la información por períodos, con el fin de analizar los momentos de mayor interés.

El análisis de series de tiempo parte de la premisa de que los datos recopilados son producto del comportamiento conjunto de tres componentes que es posible descomponer y observar: la tendencia, la estacionalidad y un componente aleatorio.

El componente aleatorio no responde a comportamientos específicos pues, como su nombre lo indica, es aleatorio o fortuito. En cambio, los otros dos brindan información más precisa sobre el comportamiento de la serie y se pueden analizar por separado. La tendencia mide el cambio de la media de los datos en el largo plazo y la estacionalidad muestra la periodicidad en los cambios, generalmente en un año, de manera que, en este caso, es posible saber cuáles son los meses con mayor protesta ciudadana y los que típicamente son más calmos.

Para llevar a cabo este análisis se usó el paquete estadístico R, con las siguientes librerías de ese *software*: "openxlsx", "xts", "itsmr", "forecast" y "dygraphs". La investigación se desarrolló en dos niveles, a saber:

a) **Descomposición y análisis.** Se descompuso la serie de tiempo para obtener por separado la estacionalidad y la tendencia de las acciones colectivas. Para ello se usó la base de datos completa (enero de 1992 a marzo de 2016). De ahí se extrajeron los factores de tendencia y estacionalidad y se graficaron para toda la serie. Luego se realizó un análisis de tendencia para cada uno de los actores registrados y se graficaron los que mostraron cambios más importantes, específicamente los trabajadores públicos, por un lado, y la ciudadanía, por el otro. Este último grupo se construyó mediante la unión de los actores que en la base de datos se denominan grupos de vecinos, madres y padres de familia, jóvenes y grupos de ciudadanos.

b) **Predicción y comparación.** Se realizó una predicción del número de las acciones colectivas para compararla con la cantidad real de movilizaciones sociales ocurridas en tres períodos específicos. Finalmente se hizo una proyección para el año 2017. Los períodos de los cuatro escenarios se detallan en el cuadro 7.11).

Aportes metodológicos en materia de descontento ciudadano

Encuesta de cultura política Barómetro de las Américas Costa Rica 2015

En 2015 se efectuó una ronda más de la encuesta de cultura política "Barómetro de las Américas" del Latin American Public Opinion Project (Lapop, por su sigla en inglés). Esta encuesta, cuyo origen se remonta a 1978, tiene ya una larga tradición en Costa Rica, donde es

CUADRO 7.11

Períodos utilizados en la predicción de las acciones colectivas

Datos para la proyección	Período proyectado	Comparación con los datos reales
Enero 2005 a diciembre 2012	Enero a diciembre 2013	Enero a diciembre 2013
Enero 2005 a diciembre 2013	Enero a diciembre 2014	Enero a diciembre 2014
Enero 2005 a diciembre 2014	Enero a diciembre 2015	Enero a diciembre 2015
Enero 2005 a marzo 2016	Abril 2016 a marzo 2017	Enero a marzo 2016

desarrollada en conjunto por la Universidad de Vanderbilt, de Estados Unidos, y el Programa Estado de la Nación. Los aportes de esta iniciativa han permitido documentar los cambios en las percepciones sobre la legitimidad del sistema político, la tolerancia y la confianza en instituciones clave como los partidos políticos, los tribunales de justicia y la Asamblea Legislativa. Asimismo, han llevado a identificar una preocupante tendencia de largo plazo, de caída en el apoyo ciudadano a la democracia. En este apartado se describen las características de la encuesta, que sirvió como principal insumo para los análisis presentados en el capítulo 6 de este Informe.

Descripción técnica del diseño muestral

Universo

El universo de la encuesta es todo el territorio continental de Costa Rica.

Población

Las unidades objeto del estudio son personas de 18 años o más, costarricenses o residentes permanentes en el país.

Unidad de observación

La unidad estadística de observación es el hogar. Toda persona entrevistada debe pertenecer a un solo hogar. En este estudio no se hace distinción entre hogares y viviendas, es decir, se considera que todo hogar habita una vivienda. Aunque ésta puede ser compartida con otros hogares, esa situación es rara en Costa Rica. La vivienda es una unidad de fácil identificación en el terreno, con relativa permanencia en el tiempo, característica que motiva su selección.

Consideraciones para el muestreo

La selección de métodos de muestreo tuvo en cuenta las consideraciones que se detallan a continuación:

a) Obtener muestras representativas para los siguientes estratos:

- Total del país
- Tamaño del municipio
  - Municipalidades con menos de 25.000 habitantes
  - Municipalidades con población de entre 25.000 y 90.000 habitantes
  - Municipalidades con más de 90.000 habitantes
- Estratos de la primera etapa
  - Área Metropolitana de San José (AMSJ)
  - Resto del Valle Central (VC)
  - Fuera del Valle Central (FVC)
- Estratos de la segunda etapa
  - Área urbana
  - Área rural

b) Efectuar cálculos de los errores de muestreo que corresponden a estos estratos.

c) Facilitar la operatividad de la encuesta.

d) Afijación óptima, que permita un equilibrio razonable entre presupuesto, tamaño de la muestra y precisión de los resultados.

e) Utilizar el mejor y más actualizado marco de muestreo disponible.

f) Muestra autoponderada.

g) Tamaño muestral de 720 entrevistas.

h) Muestra en 30 cantones (municipios) con 24 entrevistas en cada uno, salvo en el caso de San José, que cuenta dos veces) con 48, para un total de 720.

i) Conglomerados compuestos de "segmentos censales" (definidos por el INEC) de 6 entrevistas por segmento (6 x 4 = 24 por cantón).

A partir de estos antecedentes, el método utilizado correspondió a un sistema de muestreo probabilístico en todas sus etapas, estratificado, multietápico, por conglomerados, con selección aleatoria de unidades en cada etapa, incluyendo la escogencia final del adulto por entrevistar en el hogar de muestra.

El muestreo es estratificado por tamaño de las municipalidades, regiones (AMSJ, VC y FVC) y áreas (urbana y rural) y es multietápico porque parte de la selección de unidades primarias de muestreo (UPM, cantones), seguidas de unidades secundarias en cada UPM, conformadas por segmentos censales estratificados en las áreas y unidades finales de muestreo compuestas por conglomerados (segmentos compactos) de tamaño 6. En cada vivienda se selecciona y entrevista a una y solo a una persona en edad de votar, mediante un proceso aleatorio (fecha de cumpleaños más cercana a la entrevista). Como norma de selección probabilística, no se admite sustitución ni reemplazo de las unidades.

La muestra considera la asignación de tamaños que aseguran la consistencia, suficiencia y eficiencia para cada estrato y en términos agregados totales. La muestra es autoponderada a nivel nacional y dentro de cada uno de los estratos. En estos últimos la selección se realiza con probabilidad proporcional al tamaño de cada dominio.

Marco muestral

El marco muestral está constituido por el inventario cartográfico del Censo de Población y Vivienda de 2011, que identifica los segmentos censales (grupos de alrededor de sesenta viviendas definidos con propósitos de enumeración) y las viviendas que los integran. En una visita

preliminar a los segmentos seleccionados, se efectuaron actualizaciones cuando se identificaron cambios importantes con respecto al mapa usado en el Censo. La cartografía, proporcionada por el INEC, está vinculada a un archivo de Google Earth, en el cual se pueden ver las calles y casas dentro de cada segmento, con sus respectivas coordenadas geográficas.

Tamaño de la muestra

Para determinar el tamaño de muestra se utilizó un procedimiento de muestreo por conglomerados finales (DEF, por sus siglas en inglés derivadas de *design effect factor*) de tamaño 6 en áreas urbanas y rurales. Esta última es la variable explicativa del diseño y la función de la variabilidad (Kish, 1987). El efecto diseño resultante del DEF se estimó de manera preliminar en 1.1, en promedio. Esa estimación y la de los errores muestrales se basó en los datos de la encuesta. El DEF mide la relación de varianzas del diseño de muestreo empleado, por conglomerados, con respecto a un muestreo simple aleatorio. Este valor fluctúa entre 1,0 y 2,0, y tiende a ser menor conforme cuanto más pequeños son el conglomerado y la varianza real de la muestra estratificada.

Selección de la muestra

En una primera etapa se seleccionaron las unidades primarias de muestreo (UPM), en este caso los 81 cantones de Costa Rica, dentro de cada uno de los estratos (AMSJ, VC y FVC) y según tamaño de los cantones, con afijación proporcional al tamaño del estrato. La selección de los cantones dentro de cada estrato se efectuó con probabilidad proporcional al tamaño (PPT) del cantón (personas de 18 y más años que no residen en viviendas colectivas), de manera sistemática y con arranque aleatorio. El cuadro 7.12 muestra los sitios en las tres grandes regiones. El cantón San José, que tiene mayor población, se repite dos veces en la muestra. Esta última incluye, por tanto, 29 cantones o municipalidades, con 24 entrevistas cada uno, excepto San José, que tiene 48.

En una segunda etapa se seleccionaron los segmentos censales dentro de cada cantón, previa estratificación urbano-rural, con afijación proporcional al tamaño del estrato en el cantón. La selección fue también con PPT de la población votante del segmento, de manera sistemática y arranque aleatorio dentro de cada estrato, urbano o rural. Según el Censo de 2011, cada segmento tiene en promedio 125 individuos de interés con desviación estándar de 50. Para fines censales, el país está dividido en 17.200 segmentos de aproximadamente 60 viviendas cada uno. El número de segmentos en cada cantón-estrato se estableció considerando el requisito de formar conglomerados de tamaño 6, tanto en el área urbana como para la rural.

En una tercera etapa se dividió cada segmento en segmentos más compactos, cada uno con el número deseado de viviendas. Posteriormente, de manera aleatoria se seleccionó un compacto por segmento.

En total la muestra de 2015 estuvo constituida por 238 puntos: 70 en el AMSJ, 79 en otras zonas urbanas y 89 en el área rural, distribuidos en 29 cantones. El cuadro 7.13 presenta el número de segmentos que finalmente resultaron seleccionados por estrato y compara la distribución de las entrevistas por estratos en la muestra con la del Censo. Se observa que la muestra reprodujo bien la composición de la población por estratos.

En 2015 se aplicaron dos cambios con respecto a estudios anteriores:

- Se usaron los segmentos censales de la muestra de 2014, cuyo número fue de 194.
- De estos 194, se seleccionaron (al azar) únicamente 4 segmentos por cantón. Entonces, se tiene 30 cantones (contando dos veces a San José) x 4 segmentos, para un total de 120. En cada segmento se entrevistó a 6 personas, es decir, 120 x 6 = 720.

Selección de individuos

Se determinó que la muestra debía estar conformada en partes iguales por ambos sexos, o sea, 12 hombres y 12 mujeres. Para la localización geográfica de los segmentos seleccionados se utilizó un archivo en formato de Google Earth.

Niveles de confianza y márgenes de error

En encuestas demográficas con diseños similares, el DEF ha sido, en el peor de los casos, del orden de 1,1. Puede afirmarse que el error máximo para porcentajes en la muestra nacional es de 2,8

CUADRO 7.12

Cantones seleccionados por estrato, según Censo 2011[a/]

Área Metropolitana de San José		Resto del Valle Central		Fuera del Valle Central	
Cantón	Población	Cantón	Población	Cantón	Población
San José[b/]	288.054	Puriscal	33.004	Pérez Zeledón	134.534
San José[b/]	288.054	Santa Ana	49.123	San Carlos	163.745
Escazú	56.509	Alajuela	254.886	Sarapiquí	57.147
Aserrí	57.892	San Ramón	80.566	Carrillo	37.122
Desamparados	208.411	Grecia	76.898	La Cruz	19.181
Goicoechea	115.084	Poás	29.199	Puntarenas	115.019
Alajuelita	77.603	Cartago	147.898	Garabito	17.229
Tibás	64.842	Turrialba	69.616	Limón	94.415
Montes de Oca	49.132	Oreamuno	45.473	Pococí	125.962
		Heredia	123.616	Guácimo	41.266
		Belén	21.633		

a/ Población de 18 y más años, residente en viviendas no colectivas, según datos del Censo de 2011.

b/ El cantón de San José, que tiene una población sustancialmente mayor que el resto, se repite dos veces en la selección sistemática con arranque aleatorio y probabilidad proporcional al tamaño (PPT).

CUADRO 7.13

Distribución de la población y la muestra, por estrato. 2015

	Votantes según Censo 2011		Muestra		
Estrato	Número	Porcentaje	Segmentos	Entrevistas	Porcentaje
Área Metropolitana de San José	594.464	27,4	72	453	29,4
Resto Central urbano	493.171	22,7	50	325	21,1
Valle Central rural	360.153	16,6	38	242	15,7
Urbano no central	266.688	12,3	29	196	12,7
Rural no central	455.327	21,0	51	325	21,1
Total	2.169.803	100,0	240	1.541	100,0

puntos porcentuales, con 95% de confianza. Cuando la muestra se desagrega por estratos, este error puede llegar a 8 puntos porcentuales en el estrato más pequeño (urbano no central).

Ajuste por no cobertura y no elegibilidad

Para asegurar la eficiencia, suficiencia y precisión de la muestra se adoptó un sistema de muestreo con "ajuste por no cobertura", el cual garantiza la ejecución de la encuesta con los tamaños estimados como mínimos dentro de los niveles de confianza y de error máximo deseables. Adicionalmente, dicho mecanismo asegura la eliminación de sesgos resultantes de la sustitución o reemplazo de unidades que no pueden ser entrevistadas. El método requiere algún conocimiento de la "no cobertura" observada en estudios similares y la probable proporción de unidades elegibles en cada conglomerado.

El sistema consiste en aplicar a los tamaños de muestra (n) estimados para cada UPM un factor de no cobertura (t) y otro de no elegibilidad (e), con lo cual se calcula el tamaño operativo final de selección (n*), dado por:

$$n^* = (1 + t)\,(1 + e)\,n$$

Donde:

t = razón de no entrevista. Considera situaciones de no cobertura (no entrevista, rechazos, ausencia del adulto o imposibilidad de entrevistarlo después de la tercera visita, entre otros posibles eventos). Según las experiencias de otras encuestas, la tasa (t) varía por estrato y nivel socioeconómico del hogar. La tasa promedio para la muestra nacional se estimó en 0,20, lo cual significa que los entrevistadores recibieron listados con un número de viviendas un 20% más grande.

e = razón de no elegibles para la entrevista debido a discapacidad o a que no son ciudadanos costarricenses o residentes permanentes. La discapacidad se asumió como proporcional al número de adultos mayores de 75 años de edad en el segmento censal, con un promedio nacional de 3%. La proporción de extranjeros es sumamente variable, de 0% a 98% en los 194 segmentos, para un promedio nacional de 11%. En consecuencia, en un segmento donde alrededor de la mitad de la población es extranjera, se seleccionó el doble de viviendas.

Análisis de las entrevistas sobre descontento ciudadano

A inicios de 2015, en el marco de la preparación del cuestionario del estudio de cultura política de Lapop de ese año, el PEN llevó a cabo veinte entrevistas en profundidad a ciudadanos de distintas características sociodemográficas y lugares de residencia. Para la conversación se construyó una guía semiestructurada, pues la intención era profundizar en la opinión de los consultados sobre la situación del país en general. Para el presente capítulo se echó mano de ese material de investigación y, de una manera novedosa, se exploró la manera en que las y los costarricenses se refieren al descontento en su vivencia cotidiana.

Para la interpretación de las entrevistas en profundidad se empleó una técnica novedosa de análisis de redes aplicada al texto de la transcripción.

El estudio se realizó en dos etapas, denominadas recuperación de la información (RI) y análisis de redes sociales (ARS). La primera consistió en aplicar a las transcripciones una herramienta llamada "analizador léxico", lo que permitió generar, para cada entrevista, un texto base para evaluación posterior. Ese texto contiene únicamente palabras en minúscula y sin tildes; se descartaron los valores numéricos, los signos de puntuación, los artículos y las preposiciones.

El siguiente paso fue depurar el texto base, mediante la eliminación de todas las palabras que no poseían carga semántica relevante para el análisis. Estos vocablos se denominan *stopwords,* o palabras vacías. Entre las palabras suprimidas se pueden mencionar los verbos "ser" y "estar" y los que se usan para expresar criterios personales, como "creer" y "pensar", así como las muletillas propias de entrevistas no editadas, como "por ejemplo", "tal vez", "por dicha", "verdad", "mmm", "ajá", entre otras.

Luego se procedió a realizar lo que se conoce como "derivación de los términos", que consiste en buscar conceptos asociados según sus raíces, con el fin de agrupar palabras bajo una misma serie de caracteres; por ejemplo, la raíz "educ" agrupa a educación, educativa, educadoras, etc. Para esta tarea se implementó el algoritmo de *stemming* de Martin Porter (1980), uno de los más conocidos y usados para estos efectos. Al finalizar este ejercicio se contaba con una lista de menciones frecuencias de los términos con menciones de un documento, que fue el insumo para la segunda fase del proceso.

En la etapa de ARS, la finalidad es extraer una red conceptual de los térmi-

nos obtenidos en la fase previa. Con ese propósito se siguió un enfoque similar al de Paranyushkin (2011), según el cual la relación entre las palabras está dada por su proximidad en el texto. Así pues, se recorren todos los vocablos del texto base y a cada uno se le atribuye una relación con al menos dos términos próximos. La proximidad, es decir, la cantidad de palabras que es necesario recorrer antes de asignar una relación, está determinada por el tamaño del "salto" o distancia deseada entre las palabras, según criterio experto. Para esta investigación se decidió que el salto fuera de tres términos, valor que mostró una mayor expresividad y poder interpretativo del objeto de interés en las primeras pruebas. Esto implica que para cada uno de los vocablos de la lista (los cuales pueden repetirse) se establece una relación con los dos términos que se encuentren a tres palabras de distancia. Si una relación entre términos se repite, se aumenta su peso, como un indicador de fuerza de la relación.

Una pequeña modificación fue el uso de descriptores, cuya ocurrencia en el texto se conocía *a priori*. Los descriptores son palabras "clave" para categorías la creación de nodos y conformación de relaciones. Algunos de estos descriptores fueron: país, gobierno, política, diputados, economía, educación y personas.

Al concluir las dos fases descritas se contaba con dos archivos, uno que representa los nodos de la red (los términos) y otro que describe las relaciones entre ellos. Para visualizar los resultados se usó el programa Gephi (Gephi.org, 2016), mediante el cual se determinó la relevancia de cada nodo aplicando el algoritmo PageRank (Page et al, 1999). Con este último se hace un balance entre la cantidad de relaciones que tiene el nodo y el peso de cada una de ellas, de manera que se obtiene un indicador de la importancia relativa del nodo en la red.

Finalmente, y también para efectos de visualización de la red, se utilizó el algoritmo Fruchterman y Reinhold (1991), que organiza los nodos en función de su relevancia y, al igual que el anterior, toma en cuenta la cantidad de enlaces con los que el nodo se relaciona y el peso asociado a ellos.

La visualización se efectuó aplicando los siguientes criterios:

- Visualización del 5% de los nodos con los valores mñas altos de PageRank (nodos más relevantes).
- Los nodos se representan con una escala de amarillo-rojo, donde el amarillo corresponde a los PageRanks más bajos y el rojo a los valores más altos.
- El tamaño de los nodos se grafica de acuerdo con el valor del PageRank en una escala 10-120, es decir, el nodo con menor PageRank tiene tamaño 10 y el nodo más grande tamaño 120. Esta escala se refiere a la frecuencia con la que aparece una palabra en función de la cantidad de veces que esa misma palabra da sentido a otro término al antecederla.

Este Anexo Metodológico fue preparado por Ronald Alfaro, Karen Chacón, Steffan Gómez, María Estelí Jarquín, Pamela Jiménez, Natalia Morales y Rafael Segura.

NOTAS

1 La técnica consiste en tomar un conjunto de individuos (de una base de datos), cada uno de los cuales posee un conjunto de variables (atributos) denominado "x", y una variable (atributo) adicional que es la clase denominada "y". El objetivo de la clasificación es encontrar un modelo (una función o algoritmo) para predecir la clase a la que pertenecería cada individuo, asignación que se debe hacer con la mayor precisión posible. Se usa un conjunto de prueba (o tabla de testing) para determinar la precisión del modelo.

2 Para ordenar los proyectos según actividades productivas, la Setena utiliza una versión de la Clasificación Industrial Internacional Uniforme (CIIU) diseñada por la Sección de Estadísticas de la ONU. Es importante señalar que esa herramienta presenta una limitación, ya que solo categoriza la actividad que está siendo descrita, lo que puede no reflejar adecuadamente su función o propósito principal. Por ejemplo, las actividades de construcción ciertamente refieren a proyectos de edificación de estructuras, pero la categoría no distingue los objetivos finales de esa construcción con respecto a la CIIU (bodegas industriales, edificios de apartamentos, centros comerciales, etc.).

3 El FEAP permite identificar los efectos positivos o negativos latentes de una actividad, obra o proyecto sobre el ambiente, así como definir los instrumentos de manejo ambiental que el desarrollador debe presentar a la Setena.

4 Antes de la promulgación de la Ley Orgánica del Ambiente, nº 7554.